НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ»
ФАКУЛЬТЕТ ЕЛЕКТРОНІКИ
КАФЕДРА АКУСТИКИ ТА АКУСТОЕЛЕКТРОНІКИ

«На правах рукопису»
УДК _____

«До захисту допущено»
Завідувач кафедри
_____ _____
(підпис) (ініціали, прізвище)
“_–__” _____20__ р.

Магістерська дисертація
на здобуття ступеня магістра

зі спеціальності 171 «Електроніка»
(код і назва)

на тему: «Системи автоматичного розпізнавання зашумленої мови»
_____
_____

Виконала: студентка 6 курсу, групи ДГ-61м
(шифр групи)
Кухарічева Катерина Андріївна _____ _____
(прізвище, ім’я, по батькові) (підпис)

Керівник д.т.н., професор Продеус А. М. _____ _____
(посада, науковий ступінь, вчене звання, прізвище та ініціали) (підпис)

Консультант _____ _____ _____
(назва розділу) (посада, вчене звання, науковий ступінь, прізвище, ініціали) (підпис)

Рецензент _____ _____ _____
(посада, науковий ступінь, вчене звання, науковий ступінь, прізвище та ініціали) (підпис)

Засвідчую, що у цій дипломній роботі немає запозичень з праць інших авторів без відповідних посилань.
Студент _____
(підпис)

Київ – 2018 року

Завдання на магістерську дисертацію

Національний технічний університет України
«Київський політехнічний інститут»

Факультет Електроніки _____
(повна назва)

Кафедра Акустики та Акустоелектроніки _____
(повна назва)

Рівень вищої освіти – другий (магістерський)

Спеціальність 171 Електроніка_____
(код і назва)

ЗАТВЕРДЖУЮ
Завідувач кафедри
_____ _____
(підпис) (ініціали, прізвище)
«___»_____20__ р.

ЗАВДАННЯ
на магістерську дисертацію студенту

Кухарічевій Катерині Андріївні _____
(прізвище, ім'я, по батькові)

1. Тема дисертації «Системи автоматичного розпізнавання зашумленої мови» _____

керівник роботи Продеус Аркадій Миколайович, д.т.н., професор_____,
(прізвище, ім'я, по батькові, науковий ступінь, вчене звання)

затверджені наказом по університету від «___»_____ 20__ р. №_____

2. Термін подання студентом роботи _____

3. Об'єкт дослідження: системи автоматичного розпізнавання мови_____

4. Предмет дослідження: підвищення точності розпізнавання шляхом навчання на зашумлених сигналах_____

5. Перелік завдань, які необхідно розробити: 1) огляд літературних джерел; 2) організація та проведення експериментальних досліджень; 3) аналіз отриманих результатів; 4) оформлення дисертаційної роботи_____

6. Орієнтовний перелік ілюстративного матеріалу: презентація із 12 слайдів.

7. Орієнтовний перелік публікацій: 1) 2016 IEEE 4th International Conference "Methods and Systems of Navigation and Motion Control". – 18-20 Oct 2016 // Kyiv, Ukraine;
2) Electronics and Control Systems №3 (49) 2016. – NAU, Kyiv, Ukraine. ISSN: 1990-5548
3) 2nd IEEE International Conference on Advanced Information and Communication Technologies (IEEE AICT) – 4-7 July, 2017// Lviv, Ukraine

8. Консультанти розділів дисертації*

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		завдання видав	завдання прийняв

9. Дата видачі завдання _____

Календарний план

№ з/п	Назва етапів виконання магістерської дисертації	Термін виконання етапів магістерської дисертації	Примітка
1	Огляд літератури	30 листопада 2016 р.	
2	Написання вступу	15 грудня 2016 р.	
3	Написання 1-го розділу	15 березня 2017 р.	
4	Написання 2-го розділу	30 червня 2017 р.	
5	Проведення дослідження та оформлення результатів	30 жовтня 2017 р.	
6	Написання 3-го розділу	15 січня 2018 р.	
7	Написання 4-го розділу та висновків до роботи	30 лютого 2018 р.	
8	Написання реферату та оформлення всієї роботи	1 квітня 2018	

Студент _____ К.А. Кухарічева
(підпис) (ініціали, прізвище)

Керівник роботи _____ А.М. Продеус
(підпис) (ініціали, прізвище)

* Консультантом не може бути зазначено керівника дипломної роботи.

РЕФЕРАТ

Метою даної магістерської дисертації є дослідження ефективності системи автоматичного розрізнавання мови при роботі з зашумленими сигналами. Проведено оцінювання точності роботи системи автоматичного розрізнавання мови за допомогою створених адитивних сумішей мовного сигналу та реального шуму експлуатації. Для досягнення мети роботи використано портативний програмний комплекс The hidden Markov Model Toolkit. За отриманими результатами створено рекомендації стосовного вибору способу навчання системи при моделюванні систем автоматичного розпізнавання мови.

Обсяг основного тексту – 63 сторінки. Робота містить 11 рисунків, 17 таблиць, 1 додаток та 20 бібліографічне найменування за переліком посилань.

Ключові слова: автоматичне розпізнавання мови, якість системи розпізнавання, точність розпізнавання, шуми навколишнього оточення, прихована марковська модель.

ABSTRACT

This thesis is devoted to the modeling of automatic speech recognition systems for the recognition of noised speech. For reaching the goal of the thesis The Hidden Markov Model Toolkit (The HTK) was used. The additive mixtures of speech and environmental noises was created for the accuracy assessment of automatic speech recognition system. Resting on the premises of obtained results, recommendations for the modelling of automatic speech recognition systems were formulated.

The volume of the main part is 63 pages. This thesis contains 11 figures, 17 tables, 1 appendix, and 20 references.

Keywords: automatic speech recognition, recognition system accuracy, environmental noise, noised speech, Hidden Markov Model.

ЗМІСТ

ПЕРЕЛІК СКОРОЧЕНЬ ТА УМОВНИХ ПОЗНАЧЕНЬ

АРМ – автоматичне розпізнавання мови

ЛП – лінійне прогнозування

ПММ – прихована марковська модель

ШНМ – штучна нейронна мережа

FMT – Fully-Matched Training

HMM – Hidden Markov Model

HTK – Hidden Markov Model Toolkit

MT – Multistyle Training

NMT – Noise-Matched Training

SNR – Signal-to-Noise Ratio – відношення сигнал-шум

SNRMT – Signal-to-Noise Ratio Matched Training

ВСТУП

Зі стрімким розвитком науки і техніки системи автоматичного розпізнавання мови (АРМ) також набувають все більшої значущості як органи керування технічними приладами та пристроями у широкому спектрі сфер людської діяльності: науці, медицині, криміналістиці, промисловості, транспортній галузі, військовій справі, туризмі, побуті, офісній роботі тощо. Не зважаючи на значний рівень розвитку систем АРМ, важливо вдосконалювати їх, робити зручнішими та доступнішими для користувачів з різних галузей та різним рівнем кваліфікації. Особливої актуальності це набуває у ситуаціях, коли комп'ютер отримує команди від різних дикторів (має працювати з різними тембрами), які вимовляють довільний (не вузькогалузевий) текст в довільних шумових умовах, тобто за реальних умов експлуатації пристроїв. Це може бути персональний комп'ютер, працюючий в офісі типу open-space, інформаційний термінал у фойє вокзалів або на вулицях, пристрій керування безпілотного літального апарату, що підпадає під шуми навколишнього середовища, та загалом різні пристрої, що знаходяться не в спеціальному звукоізольованому приміщенні. Виходячи з цього, важливим і актуальним є виявлення впливу шумової завади на точність автоматичного розпізнавання мовленнєвого сигналу.

Метою даної магістерської дисертації є дослідження ефективності системи АРМ при роботі з зашумленими сигналами.

Задача дисертації - пошук шляхів підвищення точності розпізнавання мови, спотвореної шумами навколишнього середовища. Способо розв'язання поставленої задачі є побудова залежностей точності розпізнавання від ступеню зашумленості мовленнєвих сигналів, що використовуються при навчанні системи АРМ та її тестуванні, а також від власне типу навчання, та надання систематизованої теоретичної інформації з даного питання.

Дана дисертація є своєрідним "місточком" до наступного етапу, що планується проробити в подальших дослідженнях – розширення обсягу

словника, переліку видів шумів навколишнього середовища та збільшення кількості дикторів. Результати роботи будуть корисними при подальших розробках і дослідженнях, а також при виборі системи АРМ та типу її навчання для оснащення приладів та пристроїв промислового, військового, медичного, наукового та побутового призначення.

РОЗДІЛ 1
ТЕОРЕТИЧНИЙ ОГЛЯД СИСТЕМ АВТОМАТИЧНОГО РОЗПІЗНАВАННЯ МОВИ

Даний розділ присвячено теоретичному огляду сучасних систем автоматичного розпізнавання мови. Наведено їх класифікацю, характеристики, короткий опис підходів та алгоритмів, на яких побудовано різні системи АРМ.

1.1. Класифікація систем автоматичного розпізнавання мови

Класифікація сучасних систем автоматичного розпізнавання мови відбувається за наступними ознаками [1, 2, 3]:

- **Тип мовлення.** Мову диктора можна умовно поділити на суцільну (неперервну) та дискретну. Неперервною мовою вважаються природньо вимовлені фрази і речення, дискретною – відокремлені, ізольовані слова, між якими спеціально робиться пауза тривалістю не менш, ніж 0,25 с. Неперервна мова також має два різновиди: «спонтанна» і «підготовлена». Спонтанна мова є звичайною усною мовою, часто забрудненою хезитаціями, діалектизмами, суржиком та простими помилками. Під підготовленою мовою мається на увазі читання заготовлено тексту диктором з правильною вимовою. Системи, що розпізнають неперервне мовлення, потребують більшої обчислювальної потужності порівняно з системами, що працюють з дискретною мовою. Третім різновидом систем є такі, що відокремлюють одне слово з інтервалу мови, навіть якщо в цьому інтервалі є декілька суцільно промовлених слів, тобто здійснюють пошук ключових слів (keyword spotting).

- **Залежність від диктора.** Системи, властивістю яких є відносна незалежність від диктора, дозволяють користувачу працювати з системою без попереднього налаштування і підвищують надійність розпізнавання після навчання. Така незалежність від диктора зазвичай досягається завдяки

збереженню звукових еталонів усіх найбільш типових голосів носіїв даної мови, що, безумовно, потребує в декілька разів більшої продуктивності та об’єму пам’яті. Дикторозалежні системи зазвичай потребують попереднього налаштування на голос диктора; таке налаштування займає від лічених хвилин (у найсучасніших системах) до декількох годин. Варто зауважити, що використання дикторозалежних систем можливе з деякою точністю і без попереднього налаштування на голос конкретного користувача. Ще одним різновидом систем є такі, що налаштовуються на голос певного користувача з плином їх використання. Вони мають дві особливості: по-перше, після налаштування на одного диктора вони перестають надійно працювати з іншими голосами, по-друге, їм потрібно знати, чи не зробив користувач помилку у вимові певного слова, інакше навчання системи буде невірним.

- **Розмір словника.** Системи АРМ можуть працювати з малими (десятки слів), середніми (тисячі слів) або великими (сотні тисяч слів) словниками. Розмір словника системи майже не пов’язаний з реальною кількістю слів, що можуть бути розпізнаними, а визначається кількістю слів, що потребуються для розпізнавання в певному конкретному стані системи. Для диктування текстів потрібні великі словники; малі призначені для надання простих команд комп’ютеру або приладу; ті системи АРМ, що враховують контекст для виявлення активного підсловника в конкретному стані, фактично працюють зі словниками середнього розміру.

- **Використаний алгоритм розпізнавання.** Після того, як мовний сигнал розбивається на певні частини, відбувається імовірнісна оцінка належності цих частин до того чи іншого елементу словника. Це здійснюється за допомогою одного з алгоритмів розпізнавання. Найбільше поширення отримали системи АРМ на базі прихованих марковських моделей (ПММ), нейронних мереж та штучного інтелекту. Детальний розгляд таких систем приведено в наступному підрозділі.

- **Тип структурної одиниці.** В якості структурних одиниць можуть виступати фрази, слова, фонеми, дифони, аллофони. Системи, що розпізнають

мову, використовуючи цілі слова чи фрази, називаються системами розпізнавання мови по шаблону. Вони в більшості є дикторонезалежними, їх створення є менш працезатратним, аніж створення систем, що працюють на базі виділення лексичних елементів мови. Системи розпізнавання мови по шаблону показують вищу точність та швидкість, проте потребують значного обсягу пам'яті (пропорційно до кількості слів у словнику) та навчання кожного слова. Алгоритми порівняння лексичних елементів доводиться застосовувати у випадку великих словників, оскільки обсяг необхідної пам'яті пропорційний кількості цих еталонних елементів (наприклад, звуків), та не залежить від обсягу словника.

- **Принцип виділення структурних одиниць.** В системах АРМ використовуються декілька підходів для виокремлення з потоку мови структурних одиниць: аналіз Фур'є, вейвлет-аналіз, кепстральний аналіз. Найбільш поширений підхід засновано короткочасному перетворенні Фур'є, яке переводить початковий сигнал з амплітудно-часового простору в частотно-часовий, тобто отримують спектр мовного сигналу, що змінюється у часі. При вейвлет-аналізі, на відміну від аналізу Фур'є (в якому передбачається розклад вихідної періодичної функції в ряд, в результаті чого вихідна функція може бути представлена у вигляді суперпозиції синусоїдальних хвиль різної частоти), також відбувається розклад вхідного сигналу в базис функцій, що характеризують як частоту, так і час. Тому за допомогою вейвлетів можна аналізувати властивості сигналу одночасно і в часовому просторі, і в частотному. Третій принцип виділення структурних одиниць засновано на кепстральному аналізі, який є оберненим перетворенням Фур'є від логарифму спектру сигнала; він є складнішим за своєю сутністю та більш трудомістким, аніж перші два описаних принципи. Детальніше кепстральний аналіз розглянуто у наступному розділі.

- **Призначення.** Призначення системи визначає необхідний рівень абстракції, на якому буде проводитись розпізнавання. Наприклад, для голосового набору мобільного телефону буде використано розпізнавання по

шаблону (слову чи фразі). Такі системи називаються командними. Системи диктування потребують більшої точності розпізнавання, тому воно відбувається на базі виділення лексичних елементів і при інтерпретації промовленої фрази система АРМ буде покладатися не тільки на те, що було промовлено в теперішній момент, а ще й на сказане до цього – враховувати контекст. Також в систему такого типу має бути вбудовано набір граматичних правил. Чим жорсткіші ці правила, тим простіше реалізувати систему розпізнавання і тим обмеженішим буде набір речень, які така система зможе розпізнати.

1.2. Підходи та методи, на яких моделюють системи автоматичного розпізнавання мови

На даний час існує багато різних типів систем АРМ, побудованих на різних методах. Виокремлюють три основні підходи до машинного розпізнавання мови [4]:

1. Акусто-фонетичний підхід;
2. Розпізнавання образів;
3. Штучний інтелект.

Окремо виносять штучні нейронні мережі (ШНМ) як метод, що може бути впроваджено до кожного з зазначених підходів.

1.2.1. Акусто-фонетичний підхід до задачі розпізнавання мови

Акусто-фонетичний підхід базується на теорії акустичної фонетики, в якій передбачається, що в усному мовленні існує кінцева кількість фонетичних одиниць, що різняться одна від одної, які можуть бути в достатній мірі охарактеризовані набором властивостей, наявних у мовному сигналі чи його спектрі за певний проміжок часу. Незважаючи на те, що акустичні властивості фонетичних одиниць є вкрай мінливими, вважається що ця мінливість підкоряється певним чітким правилам, які можуть бути досить просто вивчені та застосовні до вирішення практичних завдань.

На рис.1.2.1.1 наведена схема акусто-фонетичного методу розпізнавання мови. Першим кроком є система аналізу мови, що проводить спектральне представлення характеристик мовного сигналу, змінного у часі. Найбільш поширеними техніками спектрального аналізу є ряд методів з використанням гребінки (набору) фільтрів та ряд методів кодування мовного сигналу за допомогою лінійного прогнозування. Обидва ці методи здійснюють спектральний опис мовного сигналу у часі.

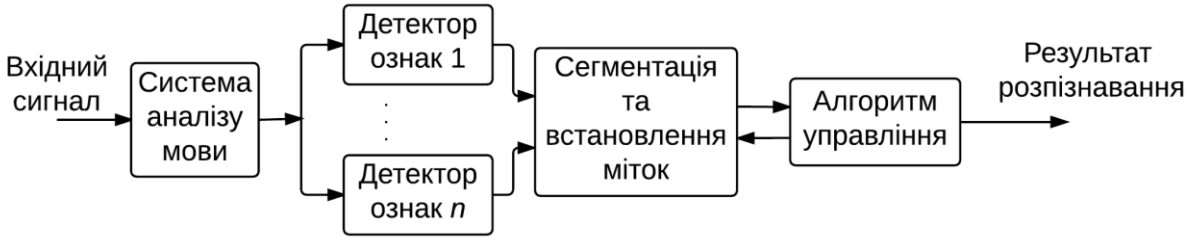

Рис. 1.2.1.1. Схема акусто-фонетичного методу

Наступним кроком акусто-фонетичного методу є виявлення ознак. Сутність цього етапу полягає в тому, щоб перевести спектральні значення, отримані у попередньому етапі, у набір рис, що описують загальні акустичні властивості різних фонетичних одиниць, до яких відносять назальність, фрикативність, розташування формант, розділ на глухі та дзвінкі звуки та співвідношення енергій низько- та високочастотних складових сигналу. Це здійснюється шляхом використання набору детекторів, що включені паралельно, які приймають рішення стосовно відсутності або наявності (та міри наявності) певної акустичної ознаки. Алгоритми, що використовуються для детектування, досить різні: це можуть бути надскладні послідовності операцій, що потребують значних обчислювальних потужностей, або тривіальні процедури оцінки.

Третім етапом є сегментація та встановлення міток, під час якого система намагається знайти стабільні ділянки (на яких відбуваються дуже малі зміни ознак), відокремити їх та встановити на них мітки, які обираються виходячи з належності проаналізованих ділянок до певної ознаки. Цей крок і

є сутністю акусто-фонетичного методу розпізнавання і є найбільш складним у виконанні, тому широко використовуються різні алгоритми управління для обмеження кількості точок розділу (сегментації) та ймовірностей встановлення міток, щоб зменшити інтервал пошуку та значно підвищити продуктивність системи.

Результатом етапу сегментації та встановлення міток зазвичай є побудова решітки фонем (рис.1.2.1.2), виходячи з якої система визначає найбільш підходящі слова або послідовність слів. Тож, на виході системи АРМ ми отримуємо слово або послідовність слів, що найкраще відповідають послідовності фонетичних одиниць фонетичної решітки.

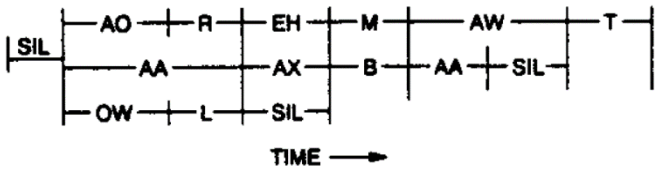

Рис. 1.2.1.2. Фонетична решітка фрази "all about" (англ. мова)

Слід зазначити, що акусто-фонетичний метод має ряд недоліків, які призводять до невдач у практичному застосуванні таких систем розпізнавання. Ці недоліки полягають у наступному:

1. Метод потребує глибоких, грунтовних знань акустичних властивостей фонетичних одиниць. На даний час ці знання є як мінімум неповними, що приводить до можливості застосування акусто-фонетичного підходу лише у простих випадках розпізнавання (наприклад, розпізнавання голосних звуків).

2. Вибір ознак в більшій мірі зроблено з ситуативних міркувань: для багатьох систем цей вибір ознак базується на інтуїції, є не дуже добре продуманим і не є оптимальним з точки зору теоретичного обґрунтування.

3. Архітектура класифікаторів також не є оптимальною. Методи, що використовують для цієї задачі, першопочатково були розроблені для побудови бінарного дерева рішень. За останні роки алгоритми побудови бінарного дерева рішень отримали певний розвиток і дерева рішень стали

менш вразливими до помилок, проте і досі їх застосування ставиться під сумнів через неоптимальність вибору ознак.

4. Поки що не існує добре обґрунтованих автоматичних процедур налагоджування методу на реальних мовних сигналах. Крім того, на сьогодення навіть не існує ідеального шляху встановлення міток на сигналах, що використовуються для тренування системи, в тому вигляді, яка б задовольнила значну кількість спеціалістів в області лінгвістики [4].

1.2.2. Штучний інтелект у задачах автоматичного розпізнавання мови

Основна ідея підходу з точки зору штучного інтелекту полягає у поєднанні знань з різних джерел і областей. У роботі [4] пропонуються такі джерела:

- Акустика – знання того, як і який звук промовлено, спираючись на спектральну інформацію про наявність чи відсутність певних акустичних ознак;
- Лексичні знання – знання того, як саме звуки пов’язуються у слова (або, відповідно, як проходить розбиття слів на окремі звуки);
- Синтаксичні знання – інформація про те, які саме комбінації слів можуть сформувати граматично вірні послідновності слів – фрази або речення;
- Семантичні знання – розуміння простору задачі, щоб визначити, чи задовольняє речення чи фраза задачі, що вирішується, та чи вони узгоджуються з попередньо розпізнаними реченнями;
- Прагматичні знання – знання, що надають здатність приймати рішення про значення неоднозначних слів та конструкцій, базуючись на тому, в якому контексті ці слова зазвичай вживаються.

Впровадження цих знань дозволяє вирішити проблеми врахування контексту, підвищення надійності при розпізнаванні суцільної мови, зниження кількості помилок завдяки здатності виправляти слова, якщо є очевидне «випадання» з загального змісту фраз чи речень.

Існує декілька способів інтегрування вказаних знань у алгоритм розпізнавання. Найбільш часто застосовується спосіб, в якому низькорівнева обробка сигналу (визначення ознак, фонетичне декодування) передує високорівневій обробці (лексичному декодуванню, мовній моделі) послідовно, щоб якнайменше впливати на кожний етап розпізнавання. У другому способі мовна модель генерує гіпотези про розпізнане слово, що відзначаються протягом усього мовного сигналу, і на основі цих гіпотез будуються семантично та синтаксично вірні речення. Третім способом є підхід на основі так званої концепції дошки оголошень, в якому всі використовувані додаткові джерела інформації вважаються незалежними, а метод висування та перевірки гіпотез є основною по'єднувальною ланкою між цими джерелами [4].

1.2.3. Штучні нейронні мережі та їх застосування в розпізнаванні мови

Як зазначено вище, застосування штучного інтелекту потребує підключення різних джерел знань. Тож, дві ключові концепції штучного інтелекту - це автоматичне набуття знань (навчання) та адаптація (навчання під час використання). Одним із способів впровадження цих концепцій є використання нейронних мереж.

Основна ідея штучних нейронних мереж, що використовують у системах АРМ, заснована на моделюванні організації та функціонування біологічних нейронних мереж – мереж нервових клітин людини, що відповідають за сприйняття мови. Вхідний сигнал аналізується «моделлю вуха», що надає спектральну інформацію про сигнал, яка зберігається у довготривалій та короткотривалій пам'яті і є доступною для різних детекторів ознак. Після декількох етапів уточнення ознак на виході системи отримуємо інтерпретацію вхідної інформації.

Нейронні мережі, незважаючи на складність їх побудови, знайшли широке розповсюдження при вирішення різних задач, в тому числі і розпізнавання мови. Це зумовлено наступними перевагами. По-перше, через

те, що нейронна мережа є високопаралельною структурою, що складається з простих ідентичних обчислювальних одиниць, вона є кращою для проведення великої кількості паралельних обчислень. По-друге, через особливості побудови, така система є надзвичайно стійкою до шумів чи дефектів всередині структури. По-третє, вагові коефіцієнти, що використовуються при побудові зв'язків між нейронами, не обов'язково мають бути точно визначені на самому початку роботи; вони можуть бути підлаштовані в процесі роботи для покращення роботи. Це і є концепцію адаптивного навчання, «рідною» для нейронний мереж, що закладена в основі штучного інтелекту.

Існує три стандартних топології нейронних мереж: одно- та багаторівневі перцептрони, рекурентні мережі, або мережі Хопфілда, та самоорганізаційні мережі Кохонена. Зазначають, що такі системи створені для роботи зі статичними моделями (наприклад, для розпізнавання зображень), проте мова за свою сутністю – процес динамічний. Тому у задачах розпізнавання мови використовуються різні модифікації базових алгоритмів, серед яких значне поширення отримала нейронна мережа з часовою затримкою [4], мережі з великою кількістю прихованих шарів – глибокі нейронні мережі [5], та гібридні моделі – поєднання ШНМ та прихованих марковських моделей [6]. Такі мережі є досить ефективними, проте складними у побудові та потребують великих обчислювальних потужностей.

1.2.4. Підхід до задачі розпізнавання мови з точки зору розпізнавання образів

В цьому методі мовні моделі (мовні образи, speech pattern) використовуються безпосередньо, без детального визначення ознак (у акусто-фонетичному сенсі) та сегментації.

У загальному вигляді метод складається з двох основних частин: навчання та тестування. Під час навчання система отримує «знання» про промовлені слова у вигляді патернів, під час тестування проводиться порівняння нового зразка з існуючими патернами. Головна ідея полягає в тому, що якщо в навчальній вибірці існує достатня кількість зразків мови, що

будуть розпізнані під час навчання, система взмозі адекватно описати акустичні властивості цих зразків і використати ці дані для подальшого порівняння.

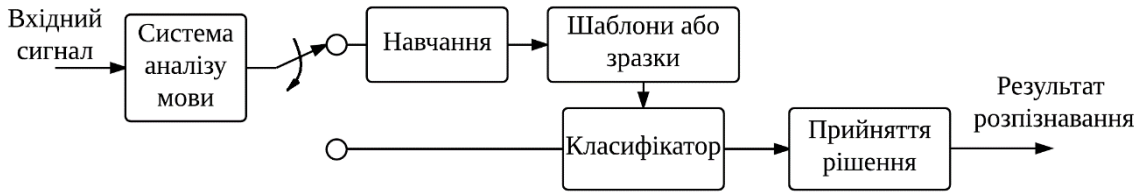

Рис. 1.2.4.1. Схема методу розпізнавання образів

Більш детальна блок-схема методу в канонічній формі наведена на рис.1.2.4.1. Концепція розпізнавання образів має в собі чотири етапи:

1. Вимірювання ознак (feature measurement). На цьому етапі проводиться обробка вхідного сигналу для того, щоб створити тестовий патерн. Для мовних сигналів використовуються алгоритми спектрального аналізу, такі як гребінка фільтрів, кодування мовного сигналу за допомогою лінійного прогнозування та дискретне перетворення Фур'є.

2. Навчання (pattern training), під час якого один чи більше тестових патернів звуків мови одного й того самого типу використовуються для створення зразкового патерну особливостей цього типу. Результуючий патерн, який називають еталонним зразком (reference pattern), може бути зразком чи шаблоном, отриманим шляхом усереднення, або моделлю, що характеризує статистичні дані ознак еталонного зразка.

3. Класифікація. На цьому етапі новий, невідомий тестовий зразок порівнюється з вже існуючими еталонними зразками та обчислюється ступінь схожості або відмінності між тестовим зразком та кожним з існуючих еталонних зразків. Для порівняння мовних зразків (що складаються з послідовності спектральних векторів) необхідно отримати міру локальної відстані, яка визначається як спектральна «відстань» між двома визначеними спектральними векторами, та провести процедуру синхронізації часу, яку ще часто називають алгоритмом динамічної трансформації шкали часу (dynamic

time warping algorithm [7]), які компенсують різницю у шкалах часу двох різних зразків.

4. Прийняття рішення (decision logic). На цьому етапі власне обчислена подібність тестового зразка до еталонних використовується для прийняття рішення: з яким з еталонних зразків найбільш добре співставляється незнайомий тестовий патерн.

Широковживаними системами АРМ, що базуються на методі розпізнавання образів, є системи, в яких патерни створюються на основі прихованих марковських моделей (ПММ). Це статистична модель, основоположним допущенням якої є те, що мовний сигнал може бути охарактеризований як параметричний випадковий процес, і що параметри цього стохастичного процесу можуть бути визначені досить точно. ПММ забезпечує високонадійне розпізнавання мови для широкого кола задач та добре інтегрується в системи, що враховують також синтаксичні та семантичні задачі.

Повертаючись до загального підходу розпізнавання образів, слід вказати на таку особливість: ефективність систем АРМ, на ньому побудованих, залежить від чисельності навчальної вибірки – тих даних, що використовуються для створення патернів для звуків певного типу; чим більша навчальна вибірка, тим вища точність розпізнавання для практично будь-якої задачі.

Загалом, метод має наступні переваги та недоліки:

+ При моделюванні системи АРМ не використовуються знання про специфіку мови, отже, така система є відносно стійкою до використання різних словників, мов, синтаксичних та семантичних завдань.
+ Оскільки система є нечутливою до типу звуків, основні техніки є застосовними до широкого кола звуків, включаючи фонетичні одиниці, цілі слова та фрази. Тож, алгоритм, розроблений для одного типу звуків, може бути використано для різних типів звуків з незначними модифікаціями або зовсім без них.

+ Процедура впровадження синтаксичних та навіть семантичних зв’язків прямо в алгоритм розпізнавання образів, що підвищує точність розпізнавання та знизити кількість обчислень, може бути досить простою та не потребує значних змін у вже існуючому алгоритмі.
- Зразковий еталон може бути вразливим до шумових умов навколишнього середовища, у якому записується мова, оскільки вони можуть істотно впливати на її спектральні характеристики.
- Завантаженість комп’ютера при навчанні та класифікації зазвичай прямопропорційна кількості зразків у навчальних та тестувальних вибірках відповідно, тож, при численній кількості типів звуків та великому об’ємі вибірок система потребує дуже великих обчислювальних потужностей [4].

1.3. Висновки

В даному розділі розглянуто класифікацію сучасних систем АРМ та теоретичне підгрунтя методів і алгоритмів, що їх використовують для побудови систем АРМ, зокрема акусто-фонетичний підхід, задачею якого є виокремлення акустичних ознак з мовного сигналу та створення на їх основі фонетичної решітки, за якою в подальшому визначається власне слово; підхід на основі використання штучного інтелекту, який поєднує у собі знання з різних галузей, пов’язаних з мовою; штучні нейронні мережі як метод, що моделює біологічні нейронні мережі та може поєднуватись з іншими основними підходами, та метод розпізнавання образів, який потребує створення зразкових еталонів розпізнаваної фонетичної одиниці та подальшого порівняння з ними тестового патерну звуку, що розпізнається. Вибір методу залежить від того, який тип задачі вирішується (чи є система командною або призначена для диктування), також потрібно враховувати умови експлуатації: обчислювальні потужності пристроїв, на які система буде встановлена, та шумові умови у приміщенні.

РОЗДІЛ 2
СИСТЕМИ АВТОМАТИЧНОГО РОЗПІЗНАВАННЯ МОВИ НА ОСНОВІ ПРИХОВАНИХ МАРКОВСЬКИХ МОДЕЛЕЙ

Даний розділ присвячено теоретичному опису прихованих марковських моделей та принципові розпізнавання окремих слів як структурної одиниці. Додатково приділено увагу одній з основних проблем експлуатації систем АРМ – несприятливим умовам, що потенційно знижують точність розпізнавання.

2.1. Приховані марковські моделі (ПММ) як основа системи розпізнавання ізольованих слів

При розгляді систем АРМ вважають, що мовний сигнал – це деяке повідомлення, що кодується за допомогою послідовності одного чи декількох символів. Для виконання зворотньої процедури розпізнавання послідовності символів, представлених у фрагменті мови, неперервний мовний сигнал спочатку перетворюють у послідовність рівновіддалених у часі векторів дискретних параметрів (показано на рис. 2.1.1). Вважається, що така послідовність векторів параметрів формує точне представлення форми хвилі мовного сигналу, оскільки на інтервалі часу, що охоплює один вектор, мовний сигнал може вважатись стаціонарним. Роль системи розпізнавання полягає в тому, що вона повинна поставити у відповідність послідовностям векторів параметрів мови підходящі послідовності символів. При розпізнаванні окремих слів припускається, що мовний сигнал відповідає одному єдиному символу у вигляді слова, обраного зі словника.

Представимо вимовлене слово послідовністю векторів або спостережень O, що визначаються як:

$$\mathrm{O} = o_1, o_2, o\ldots, o_T \qquad (2.1)$$

де o_t – вектор параметрів мови, що спостерігають в момент часу t. Проблему розпізнавання окремих слів можна розглядати як результат обчислення

$$\arg\max_i \{P(\omega_i|O)\} \tag{2.2}$$

де ω_i - це i-те слово словника. Умовну імовірність $P(\omega_i|O)$ обраховують за формулою Байєса:

$$P(\omega_i|O) = \frac{P(O|\omega_i)P(\omega_i)}{P(O)} \tag{2.3}$$

Таким чином, при заданій апріорній імовірності $P(\omega_i)$, найбільш імовірне вимовлене слово визначається лише правдоподібністю $P(O|\omega_i)$. Через високу мірність послідовності спостережень O пряме оцінювання спільної умовної імовірності $P(o_1, o_2, \ldots|\omega_i)$ на екземплярах промовлених слів на практиці не застосовується. Проте, якщо зробити припущення про параметричну модель генерування слова, таку як Марковська модель, оцінювання по даним стає можливим, оскільки проблема оцінювання умовних щільностей $P(O|\omega_i)$ заміщується більш простою задачею оцінювання параметрів Марковської моделі.

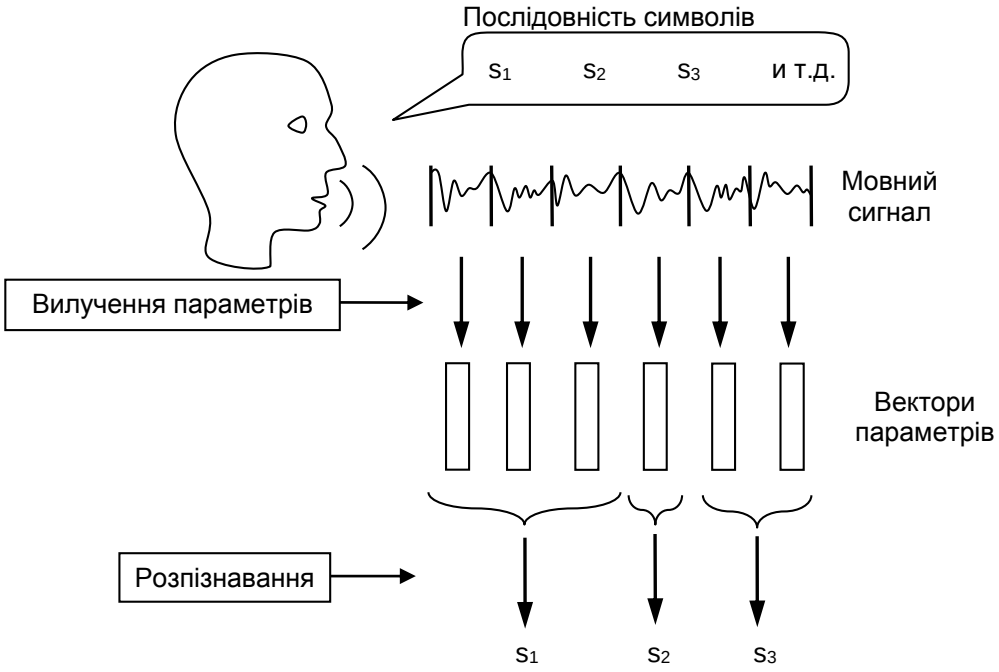

Рис. 2.1.1. Кодування/декодування повідомлення

При розпізнаванні мови, заснованому на ПММ вважається, що послідовність спостерігаємих векторів мови, що відповідають певному слову, породжена марковською моделлю, як вказано на рис. 2.1.2. Марковська

модель являє собою так званий «автомат з кінцевим числом станів», що змінює свій стан один раз за кожну одиницю часу, і в кожний момент часу t, коли модель знаходиться у стані j, вектор мови o_t генерується виходячи з щільності імовірностей $b_j(o_t)$. Окрім того, перехід від стану i до стану j також є імовірнісним та керується імовірністю a_{ij}. У представленій моделі є властивість збільшення індексу стану (або він залишається незмінним) зі збільшенням часу, отже стани переходять з одного в інший зліва направо, що утворює так звану ліво-праву модель, або модель Бакіса. Використання даної моделі дозволяє зменшити кількість можливих послідовностей станів моделі. На рис. 2.1.2 показано приклад такого процесу, де модель, що складається з шести станів, проходить через послідовність станів X=1,2,2,3,4,4,5,6, для того, щоб згенерувати послідовність від o_1 до o_6, при чому вхідний та вихідний стани ПММ не є породжуючими, що зроблено для полегшення конструювання складних моделей.

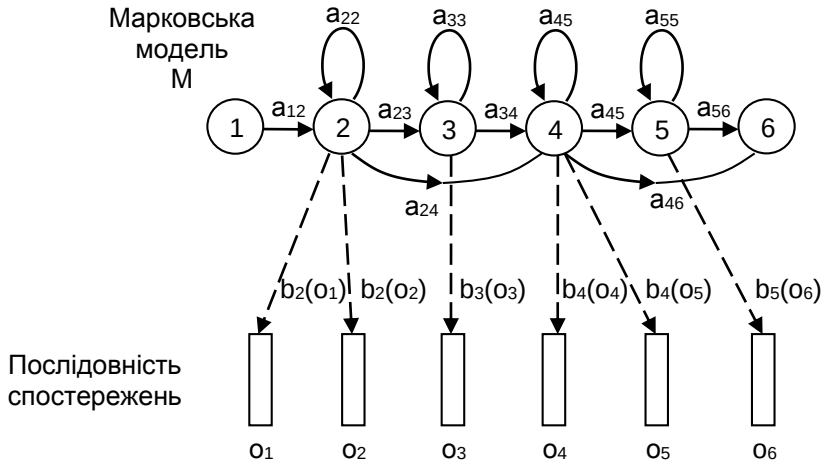

Рис. 2.1.2. Марковська модель генерування послідовності випадкових векторів

Спільна імовірність того, що O згенерована моделлю M, що проходить крізь послідовність станів X, розраховується як добуток імовірностій переходу та імовірностей генерування:

$$P(O, X|M) = a_{12}b_2(o_1)a_{22}b_2(o_2)a_{23}b_3(o_3)... \qquad (2.4)$$

Проте, на практиці відомою є лише послідовність спостережень *O*, в той час як «породжуюча» її послідовність станів *X* є прихованою від нагляду. Саме тому така модель має назву прихована марковська модель.

Оскільки *X* невідома, потрібна правдоподібність обчислюється підсумовуванням по усім можливим послідовностям станів:

$$P\left(O\middle|M\right)=\sum_{X}a_{x(0)x(1)}\prod_{t=1}^{T}b_{x(t)}\left(o_t\right)a_{x(t)x(t+1)} \tag{2.5}$$

де *x(0)* – модель вхідного стану, а $x(T+1)$ – модель вихідного стану.

У якості альтернативи до (2.5), правдоподібність може бути обчислено наближено шляхом розгляду найбільш імовірної послідовності станів:

$$P\left(O\middle|M\right)=\max_{X}\left\{a_{x(0)x(1)}\prod_{t=1}^{T}b_{x(t)}\left(o_t\right)a_{x(t)x(t+1)}\right\} \tag{2.6}$$

Для досить простого обчислення (2.5) та (2.6.) використовують прості рекурсивні формули. Слід також зазначити, що якщо співвідношення (2.2) може бути обчисленно, задача розпізнавання є вирішеною. Для заданої множини моделей, що відповідають словам, співвідношення (2.2) вирішується, використовуючи (2.3) та припускаючи, що:

$$P\left(O\middle|\omega_i\right)=P\left(O\middle|M_i\right) \tag{2.7}$$

При цьому вважається, що параметри a_{ij} та $b_j(o_t)$ відомі для кожної моделі M_i. Для заданої множини прикладів навчання, що відповідають певній моделі, параметри цієї моделі можна визначити автоматично за допомогою надійної та ефективної рекурентної процедури. Таким чином, при умові, що зібрано достатню кількість зразків кожного слова, можливо побудувати ПММ, яка неявно моделює всю множину причин мінливості, властивій реальній мові [8].

В залежності від модельованих параметрів, розподіли багатомірних щільностей імовірностей можуть бути як неперервними, так і дискретними. В даному випадку для простоти вважається, що використовуються неперервні щільності розподілів. У більшості систем, що працюють з неперервними

щільностями, розподіли описуються гауссівськими сумішами щільностей імовірності. В такому випадку формула розрахунку $b_j(o_t)$ має вигляд:

$$b_j(o_t) = \sum_{m=1}^{M} c_{jm} N\left(o_t, \mu_{jm}, \Sigma_{jm}\right), \tag{2.8}$$

де M – число компонентів, що змішуються, c_{jm} – вага m-го компоненту, а $N(o,\mu,\Sigma)$ – багатомірний гауссівский розподіл з вектором середнього значення μ та коваріаційною матрицею Σ:

$$N(o,\mu,\Sigma) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \exp\left(-\frac{1}{2}(o-\mu)' \Sigma^{-1}(o-\mu)\right), \tag{2.9}$$

де n – розмірність o [8].

2.2. Рекурентне оцінювання параметрів ПММ за алгоритмом Баума-Уелча (Baum-Welch)

Для визначення параметрів ПММ необхідно зробити припущення про те, якими вони могли б бути. Як тільки це зроблено, стає можливим віднайти більш точні параметри (в сенсі максимальної правдоподібності) за допомогою так званої рекурентної процедури Баума-Уелча. Можна припустити, що компоненти суміші є спеціальною формою стану більш низького рівня, в якому імовірності переходу є вагою суміші (див. рис. 2.3.1).

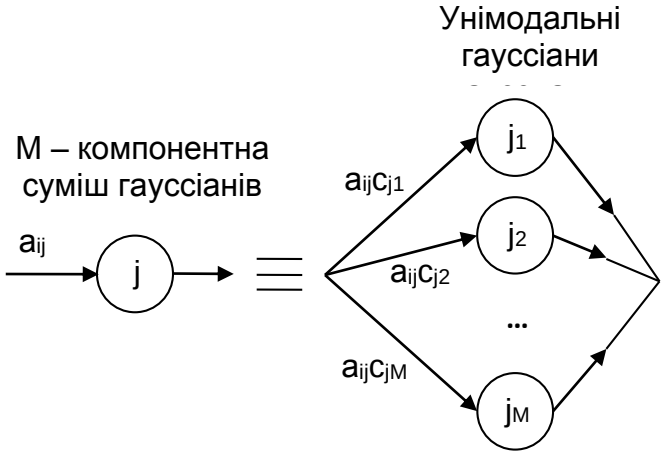

Рис.2.3.1. Представлення суміші

Таким чином, важливою задачею є оцінювання середніх та дисперсій ПММ, в якій кожен стан вихідних даних представляється єдиним гауссівським компонентом:

$$b_j(o_t) = \frac{1}{\sqrt{(2\pi)^n |\Sigma_j|}} \exp\left(-\frac{1}{2}(o_t - \mu_j)' \Sigma_j^{-1} (o_t - \mu_j) \right) \tag{2.10}$$

Якщо б ПММ могла набувати лише одного стану, оцінити максимальну правдоподібність величин μ_j та Σ_j можна було б завдяки простому усередненню:

$$\hat{\mu}_j = \frac{1}{T} \sum_{t=1}^{T} o_t \tag{2.11}$$

$$\hat{\Sigma}_j = \frac{1}{T} \sum_{t=1}^{T} (o_t - \mu_j)(o_t - \mu_j)' \tag{2.12}$$

На практиці система може набувати декількох станів і неможливо безпосередньо «прив'язати» вектори спостережень до окремих станів, оскільки базові послідовності станів невідомі. Тому здійснюють деяку наближену прив'язку векторів до станів, щоб можна було використати (2.11) та (2.12) для отримання потрібних початкових значень параметрів. Далі, використовуючи алгоритм Вітербі, знаходять найбільш правдоподібну послідовність станів, вектори спостережень заново прив'язують до станів, після чого знову використовують рівняння (2.11) та (2.12) для отримання найкращих початкових значень. Цей процес повторюється доти, доки оцінки перестають змінюватись. Оскільки повна правдоподібність кожної послідовності спостережень базується на сумуванні усіх можливих послідовностей станів, кожен вектор спостереження вносить свій вклад в розрахунки значень параметрів максимальної правдоподібності для кожного стану. Інакше кажучи, замість того, щоб прив'язувати кожен вектор спостереження до певного стану, як це було зроблено у вищевказаному припущенні, кожне спостереження прив'язується до кожного стану, пропорційно імовірності стану моделі при спостереженні цього вектора.

Отже, якщо через $L_j(t)$ позначити імовірність знаходження в стані j в момент часу t, рівняння (2.11) та (2.12) перетворюються на такі зваженими середні:

$$\hat{\mu}_j = \frac{\sum_{t=1}^{T} L_j(t) o_t}{\sum_{t=1}^{T} L_j(t)}, \tag{2.13}$$

$$\hat{\Sigma}_j = \frac{\sum_{t=1}^{T} L_j(t)(o_t - \mu_j)(o_t - \mu_j)'}{\sum_{t=1}^{T} L_j(t)}, \tag{2.14}$$

де сумування в знаменниках забезпечує необхідну нормалізацію. Рівняння (2.13) та (2.14) описують процедуру рекуррентного оцінювання Баума-Уелча для середніх та дисперсій СММ.

Імовірність стану $L_j(t)$ ефективно розраховується з використанням так званого алгоритму прямого-зворонього ходу (Forward-Backward algorithm). Нехай пряма імовірність $\alpha_j(t)$ для певної моделі M з N станами визначена у вигляді:

$$\alpha_j(t) = P(o_1,\ldots,o_t, x(t) = j) \tag{2.15}$$

Тобто $\alpha_j(t)$ – спільна імовірність спостереження перших t векторів мови для стану j в момент часу t. Ця пряма імовірність може бути ефективно розрахована за наступною рекурентною формулою:

$$\alpha_j(t) = \left[\sum_{i=2}^{N-1} \alpha_i(t-1)\alpha_{ij}\right] b_j(o_t) \tag{2.16}$$

Вигляд цієї рекурентної формули визначається тим, що імовірність знаходження в стані j в момент часу t, при спостереженнях o_t, можна вивести шляхом сумування прямих імовірностей для усіх можливих передуючих станів елементу i, зважених імовірностями переходів a_{ij}. Вказані межі сумування враховують те, що стани 1 та N не є породжуючими. Початкові умови для вищезазначеного рекурентного співвідношення мають вигляд:

$$\alpha_1(1) = 1 \tag{2.17}$$

$$\alpha_j(1) = a_{1j} b_j(o_1) \tag{2.18}$$

для $1<j<N$, а кінцеві умові задаються наступним чином:

$$\alpha_N(T) = \sum_{i=2}^{N-1} \alpha_i(T) a_{iN} \tag{2.19}$$

Слід зазначити, що з визначення $\alpha_j(t)$ випливає:

$$P(O \mid M) = \alpha_N(T) \tag{2.20}$$

Отже, обчислення прямої імовірності дозволяє отримати повну правдоподібність $P(O|M)$. Зворотня імовірність $\beta_i(t)$ визначається як:

$$\beta_j(t) = P(o_{t+1},...,o_T \mid x(t) = j, M) \tag{2.21}$$

Ця зворотня імовірність також може бути ефективно обчислена за допомогою наступного рекурентного співвідношення:

$$\beta_i(t) = \sum_{j=2}^{N-1} a_{ij} b_j(o_{t+1}) \beta_j(t+1) \tag{2.22}$$

з початковою умовою:

$$\beta_i(T) = a_{iN} \tag{2.23}$$

для $1<i<N$, та кінцевою умовою

$$\beta_1(1) = \sum_{j=2}^{N-1} a_{1j} b_j(o_1) \beta_j(1) \tag{2.24}$$

У наведених вище визначеннях пряма імовірність є імовірністю спільною, в той час як зворотня імовірність є умовною. Таке асиметричне визначення дозволяє визначити імовірність знаходження у стані як добуток цих двох імовірностей. За визначенням,

$$\alpha_j(t)\beta_j(t) = P(O, x(t) = j \mid M) \tag{2.25}$$

Звідки

$$L(t) = P\left(x(t) = j \mid M\right) = \frac{P\left(O, x(t) = j \mid M\right)}{P(O \mid M)} = \frac{1}{P}\alpha_j(t)\beta_j(t) \tag{2.26}$$

де $P = P(O \mid M)$.

Тож тепер уся інформація, що потребується для здійснення рекурентного оцінювання параметрів ПММ за допомогою алгоритма Баума-

Уелча, є відомою. Покрокове описання цього алгоритму має наступний вигляд:

1. Для кожного рекурентно оцінюємого векторного/матричного параметру слід виділити місце в пам’яті для проведення сумування в чисельнику та знаменнику виразів (2.13) та (2.14). Ці місця пам’яті являють собою накопичуючі суматори.

2. Обчислюють прямі та зворотні імовірності для усіх станів j та моментів часу t.

3. Для кожного стану j та часу t оновлюють вміст накопичуючих суматорів, використовуючи імовірність $L_j(t)$ та поточний вектор спостереження o_t.

4. Кінцеві значення накопичуючого суматора використовують для обчислення нових значень параметрів.

5. Якщо значення $P = P(O|M)$ для даної ітерації не вище такого для попередньої ітерації, виконується зупинка, в іншому випадку повторюють вищевказані кроки, використовуючи нові рекурентно оцінені значення параметрів.

З вищевказаного випливає, що параметри ПММ рекурентно оцінюються по єдиній послідовності спостереження, тобто по єдиному екземпляру промовленого слова. На практиці для отримання хороших оцінок параметра, необхідно мати багато екземплярів одного і того самого слова; тим не менш, використання багатократних послідовностей спостереження не призводять до ускладнення алгоритму, відбувається лише повторення кроків 2 та 3 для кожної нової навчальної послідовності. Також слід зазначити, що обчислення прямих та зворотніх імовірностей пов’язано з обчисленням добутку великої кількості імовірностей, це означає, що числа набувають малих порядків величин. Тож, для запобігання обчислювальних проблем, прямі-зворотні обчислення реалізуються за допомогою логарифмування [8].

2.4. Алгоритм Вітербі

Вище, при описанні основних ідей рекурентної оцінки параметрів ПММ за допомогою алгоритма Баума-Уелча, було зазначено, що ефективний рекурсивний алгоритм обчислення прямої імовірності дозволяє також обчислити і повну імовірність $P(O|M)$. Таким чином, цей алгоритм може бути застосовним для знаходження моделі, що максимізує значення $P(O|M_i)$, а, отже, може бути використаний для розпізнавання.

На практиці, проте, здійснюють розпізнавання, засноване на максимізації правдоподібності послідовності стану, оскільки це легко узагальнюється на випадок суцільної мови, що неможливе при використанні повної імовірності. Відмінність від алгоритму обчислення прямої імовірності при обчисленні правдоподібності полягає в тому, що сумування замінюється пошуком максимуму.

Нехай $\varphi_j(t)$, для даної моделі M, являє собою максимальну правдоподібність спостереження послідовності векторів мови від o_1 до o_t та перебування в стані j в момент часу t. Ц. Часткову правдоподібність можна ефективно обчислити з використанням наступного рекурентного співвідношення:

$$\varphi_j(t) = \max_i \{\varphi_i(t-1)a_{ij}\} b_j(o_t) \tag{2.27}$$

де

$$\varphi_1(1) = 1 \tag{2.28}$$

$$\varphi_j(1) = a_{1j} b_j(o_1) \tag{2.29}$$

для $1 < j < N$. Тоді максимальна правдоподібність $P(O|M)$ має вигляд:

$$\varphi_N(T) = \max_i \{\varphi_i(T)a_{iN}\} \tag{2.30}$$

Щодо рекурентного оцінювання, то пряме обчислення правдоподібності приводить до втрати значущих розрядів, тому замість цього обраховують логарифм правдоподібності. При цьому замість рівняння (2.27) маємо:

$$\psi_j(t) = \max_i \{\psi_i(t-1) + \log(a_{ij})\} + \log(b_j(o_t)) \tag{2.31}$$

Це рекурентне співвідношення є основою так званого алгоритму Вітербі. Як показано на рис. 2.4.1, цей алгоритм можна представити як виявлення найкращого шляху крізь матрицю, де вертикальний вимір – стан ПММ, а горизонтальний – фрейми мови (тобто час). Кожна велика точка на рисунку – логарифм імовірності спостереження даного фрейма в даний момент часу, а кожен відрізок між точками відповідає логарифму імовірності переходу. Логарифм імовірності будь-якого шляху обчислюється простим підсумуванням логарифмів імовірностей переходів та логарифмів вихідних імовірностей вздовж даного шляху. Шляхи проходять зліва направо, стовпчик за стовпчиком. В момент часу *t*, кожен частковий шлях $\psi_i(t\text{-}1)$ є відомим для усіх станів *i*, тому рівняння (2.31) можна використовувати для обчислення $\psi_j(t)$, продовжуючи часткові шляхи на один такт часу [8].

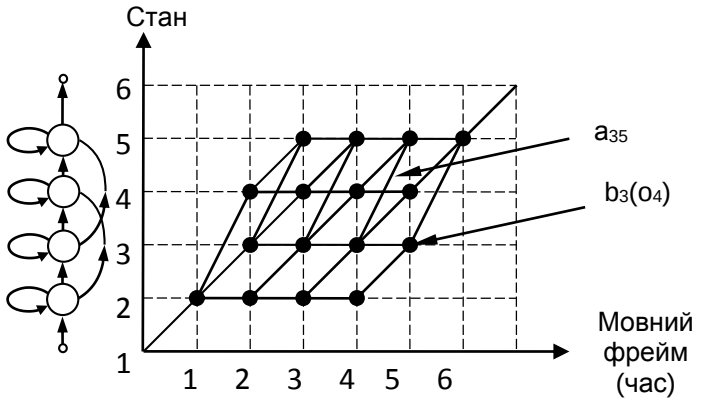

Рис.2.4.1. Алгоритм Вітербі для розпізнавання ізольованих слів

2.5. Параметризація мовного сигналу

Для розпізнавання мови, заснованого на ПММ, необхідно отримати набір векторів спостереження. В якості векторів спостереження повинні бути обрані такі параметри, за якими можна надійно відрізняти один звук мови від іншого. Форма мовного сигналу одного і того самого звука може істотно відрізнятись, тому відліки мовного сигналу безпосередньо не використовуються, частіше всього використовують різні методи спектрального аналізу, зокрема засновані на перетворенні Фур'є або методі лінійного прогнозування.

2.5.1.Аналіз на основі лінійного прогнозування

Лінейне прогнозування (лінійне передбачення, ЛП) є одним з найбільш ефективних методів, цей метод є домінуючим при оцінці основних параметрів мовного сигналу, таких як період основного тону, форманта, спектр, функція площі мовного тракту. Важливість методу зумовлена високою точністю отриманих оцінок та відносною простотою обчислень. Основний принцип методу лінійного передбачення полягає в тому, що поточний відлік мовного сигналу можна апроксимувати лінійною комбінацією передуючих відліків:

$$s(n) = -\sum_{i=1}^{N}\left[a_i + s(n-i)\right] + e(n) \qquad (2.32)$$

де N – кількість коефіцієнтів моделі, $e(n)$ – помилка передбачення.

Коефіцієнти перебдачення при цьому однозначно визначаються мінімізацією середнього квадрату $e(n)$ – різниці між відліками мовного сигналу та їх передбаченими значеннями (на кінцевому інтервалі). Основні положення методу ЛП добре узгоджуються з моделлю мовоутворення, де мовний сигнал можна представити у вигляді сигналу на виході лінійної системи зі змінними у часі параметрами, що збуджується квазіперіодичними імпульсами (в межах вокалізованого сегменту) або випадковим шумом (на невокалізованому сегменті). Метод ЛП дозволяє точно та надійно оцінити параметри цієї лінійної системи зі змінними коефіцієнтами. Функція передачі такої системи:

$$H(z) = \frac{1}{\sum_{i=0}^{p} a_i z^{-i}}, \qquad (2.33)$$

де p – число полюсів та $a_0 \equiv 1$. Застосовно до мовних сигналів існують наступні методи обчислення параметрів a_i (часто вони є рівноцінними): коваріаційний, автокорреляційний, східчастого фільтру, зворотньої фільтрації, оцінювання спектру, максимальної правдоподібності та скалярного добутку [1].

2.5.2. Кепстральні коефіцієнти

Окрім коефіцієнтів лінійного передбачення часто для розпізнавання мови використовують коефіцієнти кепстра відрізку мовного сигналу. Кепстр являє собою перетворення Фур'є від логарифма спектру сигналу. На відміну від класичного визначення кепстра, в галузі обробки мови кепстр визначається як зворотнє перетворення Фур'є від логарифму спектра потужності сигналу. В математичному вигляді це описується наступною формулою:

$$c = FT^{-1}\left\{\lg\left|FT\left\{s\right\}\right|\right\} \tag{2.34}$$

де $FT()$ – перетворення Фур'є, $FT^{-1}()$ – зворотнє перетворення Фур'є, c – кепстр, s – початковий сигнал. На відміну від комплексного спектру, у формулі (2.34) кепстр не містить інформації про фазу сигналу. Змінна, від якої залежить кепстр, має розмірність часу, проте це вже не той самий час, що і для сигнала. Щоб підкреслити, що ця змінна є в деякому сенсі частотою для кепстра, вживають поняття *quefrency*; фільтр, що здійснює операції над кепстром називаюь *lifter*.

Операція обчислення кепстра відноситься до класу гомоморфної обробки сигналу. Гомоморфні системи – це клас нелінійних систем, що підкоряються загальному принципу суперпозиції; лінійні системи є частковим випадком гомоморфної системи. В обробці мови гомоморфні системи мають наступну властивість:

$$D\left[\left[x_1(n)\right]^{\alpha}\cdot\left[x_2(n)\right]^{\beta}\right] = \alpha D\left[x_1(n)\right] + \beta\left[x_2(n)\right] \tag{2.35}$$

Цей тип суперпозиції відноситься до операцій множення та піднесення до степені. Таку узагальнену властивість суперпозиції має функція логарифма. Гомоморфні системи корисні в обробці мови, оскільки вони являють собою метод для розділення форми збуджуючого сигналу та імпульсної перехідної характеристики мовного тракту. Для розпізнавання мови цей процес є корисним з точки зору моделювання характеристик мовного тракту. Розділення двох компонент можна представити як процес деконволюції (оберенена згортка) і може бути описаний наступним чином:

$$s(n) = g(n) \otimes v(n) \tag{2.36}$$

де $g(n)$ – збуджуючий сигнал, $v(n)$ – імпульсна перехідна характеристика мовного тракту, а $\otimes$ - операція згортки. В частотному вигляді згортка представляється як:

$$S(f) = G(f) \otimes V(f) \tag{2.37}$$

Взявши комплексний логарифм від обох частин, отримаємо:

$$Log(S(f)) = Log(G(f) \otimes V(f)) = Log(G(f)) + Log(V(f)) \tag{2.38}$$

З цього моменту, в логарифмічному сенсі, збурення та характеристика мовного тратку являють собою адитивну суміш. Узявши зворотнє перетворення Фур'є цієї суміші, отримаємо кепстр сигналу, який буде являти собою суміш імпульсної характеристики та сигналу збурення, які при необхідності можна розділити методами лінійної фільтрації. Причому інформація про мовний тракт буде зосереджена в основному в області малих часів кепстра, в той час як в області більших значень часу зосереджена інформація про сигнал збурення. Замість обчислення перетворення Фур'є сигналу на практиці частіше за все використовують гребінку фільтрів. Крім того, кепстр сигналу також можна отримати з коефіцієнтів лінійного передбачення. Для цього використовується рекурсивна формула:

$$c_n = -a_n - \frac{1}{n}\sum_{i=1}^{n-1}(n-i)a_i c_{n-i} \tag{2.39}$$

При чому порядок кепстра не обов'язково дорівнює кількості коефіцієнтів LPC.

Відомо, що людське вухо має різну частотну роздільну здатність на різному діапазоні частот; інакше кажучи, перетворення сигналу в спектр проходить нелінійно, оскільки вважається, що таке нелінійне перетворення підвищує розбірливість мови. В системах розпізнавання для отримання такого перетворення використовують гребінку фільтрів різної ширини, після чого обчислюють логарифм та зворотнє перетворення Фур'є для отримання кепстра. Одна з таких нелінійних шкал, апроксимуюча шкалу частот людського слуху, називається Мел-частотною і визначається як:

$$Mel(f) = 2595\log(1 + \frac{f}{700}) \quad (2.40)$$

Характер гребінки, що використовується для цієї шкали, показано на рис.2.5.2.1 [1].

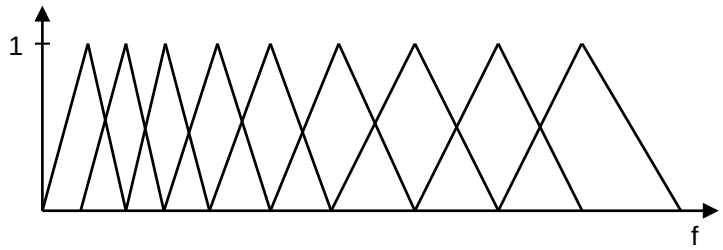

Рис. 2.5.2.1 Гребінка фільтрів мел-частотной шкали

2.5.3. Дельта-коефіцієнти

Ефективність системи АРМ може бути істотно підвищена, якщо додатково використовувати похідну по часу від основних статистичних параметрів. Ці коефіцієнти мають назву дельта коефіцієнтів першого порядку, другого порядку (прискорення) та третього порядку. Дельта коефіцієнти обчислюють наступним чином:

$$d_t = \frac{\sum_{\theta=1}^{\Theta} \theta(c_{t+\theta} - c_{t-\theta})}{2\sum_{\theta=1}^{\Theta} \theta^2}. \quad (2.41)$$

Коефіцієнти другого та третього порядків обчислюються за тією самою формулою, але від дельта коефіцієнтів попереднього порядку [1].

2.6. Чинники зниження точності розпізнавання

При моделюванні систем АРМ важливо врахувати несприятливі зовнішні умови, що значно знижують якість розпізнавання, і яких, на жаль, неможливо уникнути під час експлуатації. Виокремлюють три основні проблеми: спотворення, вплив артикуляції та шум [4].

2.6.1. Спотворення

Мовний сигнал неминуче потерпає від ряду спектральних спотворень ще до того, як буде почато процес власне розпізнавання. Приміщення, в якому встановлена система АРМ, майже стовідсотково має певний рівень реверберації, що спотворюється через прийняття прямого та відбитого сигналу через велику відстань від диктора до мікрофону [9]. Акустоелектричний перетворювач мікрофону, залежно від його типу та позиції відносно рота диктора, може також значно спотворити спектр мови. Якщо конфігурації мікрофона, що використовувався для навчання системи, відрізняються від того, що використовували для тестування, неспівпадіння спектральних спотворень стає однією з найбільших проблем, що мають великий вплив на точність розпізнавання [4].

2.6.2. Артикуляційні ефекти

На індивідуальну манеру мовлення диктора впливає багато факторів. Навіть психологічне усвідомлення факту «спілкування» з машиною може внести істотні зміни у звукові форманти та ритм промовляння. Зміни характеристик голосу під впливом навколишнього середовища (це явище також відоме як ефект Ломбарда [10]) також можуть бути значними. Під час розмови у зашумлених приміщеннях (з рівнем шуму 90 дБ SPL), часто перша форманта голосних підвищується, в той час як частота другої форманти знижується, що в результаті призводить до зсуву простору голосних. Ці зміни істотно впливають на ефективність системи: дикторозалежна система АРМ, що працює з ізольованими словами з високою точністю у випадку, коли навчання та тестування проводиться за чистих (від ефекту Ломбарда) умов, значно втрачає точність, коли тестові зразки містили в собі ефект Ломбарда, хоча й були штучно очищені від шуму.

Найбільша складність у боротьбі з артикуляційними ефектами полягає в неповному розумінні того, як надати їм кількісну оцінку. Акустичні шуми або спотворення, що виникають у звуковому тракті, зазвичай змінюються не так

швидко, як мова (у спектральному сенсі), тому в деякій мірі їх можливо виміряти або змоделювати, щоб покращити конструкцію системи. Натомість, артикуляційні зміни у вимові одних і тих самих слів є невід'ємними та контекстно залежними, а кількісні характеристики ефекту Ломбарда не є достатньо визначеними для того, щоб впроваджувати певні алгоритми для боротьби з впливом артикуляції у системи автоматичного розпізнавання мови [4].

2.6.3. Шуми навколишнього оточення

Акустичні шуми навколишнього оточення зазвичай вважаються адитивними, тобто записаний сигнал вважається сумою мовного синалу та шуму. Високі рівні шумів є однією з найбільших проблем при розпізнаванні мови, оскільки джерел шумів дуже багато та їх майже неможливо усунути. Наприклад, в офісі джерелами шуму виступає оргтехніка: принтери, персональні комп'ютери, клавіатура, телефони; невід'ємним джерелом шуму також є фонові діалоги інших людей (зазвичай рівень сягає 45 – 50 дБА). Цих шумів зазвичай достатньо, щоб спричинити значне спадання ефективності роботи системи. Другим прикладом є робота системи АРМ у автомобілі: рівень шумів двигуна, кондиціонера, шин та дороги є досить високим, при русі на високій швидкості або з відкритим вікном навіть людині іноді важко розібрати мову, не кажучи про систему розпізнавання. При чому спектр акустичного фону навколишнього середовища зазвичай не рівномірний: наприклад, низькочастотний спектр шумів в автомобілі, спричинених механічними джерелами (двигун, шини), є частотно залежним та спадає зі зростанням частоти, високочастотні шуми, спричинені аеродинамічними явищами, мають рівномірний спектр на частотах вище 1 кГц.

Інші типи шумів, такі як шум квантизації та електричні шуми, зазвичай мають низькі рівні, тож ними можна знехтувати. Тим не менш, «кліки», що виникають при включенні або відключенні певних елементів, можуть також бути заважаючим фактором [4].

2.6.4. Методи боротьби з несприятливими зовнішніми факторами

Відомо, що якщо характеристики заважаючого шуму є приблизно відомими, система АРМ, що тренована на патернах зі схожими шумовими характеристиками, загалом працює ефективніше, ніж та, що тренована на чистих сигналах. Цей факт може бути використаний і для інших випадків, наприклад тоді, коли диктор у відповідь на високий рівень шуму в приміщенні трохи змінює манеру мовлення (зазначені вище у п.2.6.2 артикуляційні ефекти). Ідея полягає в тому, щоб включити в навчальну вибірку слова з різним емоційним забарвленням і темпом промовляння; це може підвищити якість розпізнавання до 2х разів.

Проте, таке навчання системи не вирішує проблему повністю, тому застосовують ряд інших допоміжних методів: попередню обробку сигналу, за якої підвищують рівень корисного сигналу [11]; використання спеціальних технічних засобів, таких як мікрофони, оснащені фільтрами шумів, або градієнтних мікрофонів; побудову адаптивних моделей, що маскують шум; метод компенсації впливу артикуляційних ефектів [12]; метод аналізу найбільш значних спотворень [13], а також нові підходи до аналізу мовних сигналів навіть за умов наявності шумів. [4]

2.7. Висновки

Для проведення процесу розпізнавання мови необхідно поставити у відповідність мовному сигналу послідовність дискретних параметрів, що формує точне представлення форми хвилі сигналу, що розпізнається. Параметризація сигналу проводиться за методами спектрального аналізу, такими як аналіз на основі лінійного передбачення та перетворення Фур’є. Останнє дозволяє отримати мел-частотні кепстральні коефіцієнти, що безпосередньо використовуються у розпізнаванні та підвищують його ефективність, оскільки мел-частотна шкала враховує особливості частотної

вибірковості слуху людини і вважається, що використання цієї властивості підвищує розбірливість мови.

РОЗДІЛ 3
ОЦІНЮВАННЯ ТОЧНОСТІ РОЗПІЗНАВАННЯ ЗАШУМЛЕНИХ МОВНИХ СИГНАЛІВ

Даний розділ присвячено огляду мір якості систем автоматичного розпізнавання мови, показникам, що використовуються для оцінки якості, а також показникам, що використовуються при оцінюванні точності розпізнавання в системі АРМ The Hidden Markov Model Toolkit (HTK). Наведено результати експериментальних досліджень залежності ефективності роботи системи за різних варіантів навчання та тестування.

3.1. Оцінювання якості систем автоматичного розпізнавання мови

При моделюванні системи АРМ важливо надати оцінку якості її роботи. Ідея такого оцінювання є досить простою: потрібно подати на вхід системи АРМ N слів, на виході підрахувати кількість правильно розпізнаних слів. Проте, на практиці це реалізувати важко через велику кількість слів, що розпізнаються; додатковим ускладненням є наявність «вставок» - випадкових шумів, які система помилково вважає словами. Ці шуми можуть мати різне походження: це сторонні фонові шумі або ті, що створюються диктором: дихання, кашляння, ковтання та інші.

Найбільш часто при оцінюванні якості системи АРМ використовують показник Word Error Rate (WER) – відносну помилку розпізнавання [8]:

$$WER = \frac{S + D + I}{N = H + S + D} \tag{3.1}$$

та обернену йому величину – точність розпізнавання мови:

$$Acc = 1 - WER \tag{3.2}$$

де N – загальна кількість розпізнаних слів, H – кількість вірно розпізнаних слів, S – кількість замін, D – кількість видалень, I – кількість вставок. За відсутності вставок I значення показника WER належать інтервалу [0, 1], проте при дуже

великій кількості вставок показник Acc може набувати від'ємних значень, що з практичної точки зору не має сенсу.

Для усунення цього недоліку можливо не враховувати вставки у формулах для точності розпізнавання та відносній помилці розпізнавання відповідно:

$$\%Correct = \frac{H}{N} \cdot 100\% \tag{3.3}$$

$$P_e = \frac{S + D}{N} \tag{3.4}$$

проте і у цьому випадку при значній кількості вставок показники (3.3) та (3.4) не будуть вірно відображати якість роботи системи.

Для пошуку альтернативних показників до тих, що зазначені вище, слід сформувати вимоги, яким мають відповідати нові критерії [14]: по-перше, вони мають бути математично обгрунтованими, по-друге, вони повинні бути інтуітивно зрозумілими, по-третє, – відносно простими для обчислення. В роботі [14] зазначено, що таким показником може бути «узгоджений процент помилок» - Match Error Rate (MER):

$$MER = \frac{S + D + I}{N = H + S + D + I} = 1 - \frac{H}{N} \tag{3.5}$$

Окрім цього, в [14] запропоновано новий показник, що відображає «втрачену словесну інформацію» (Word Information Lost – WIL), та пов'язаний з ним показник, що описує «збережену словесну інформацію» (Word Information Preserved - WIP):

$$WIP \cong \frac{I(X,Y)}{H(Y)} \cong \frac{(H - N_1 N_2 / nN)^2}{N_1 N_2} \tag{3.6}$$

$$WIL = 1 - WIP \tag{3.7}$$

де $I(X,Y)$ - повна взаємна інформація про набори слів X та Y на вході та виході системи АРМ відповідно; $H(Y)$ - ентропія послідовності слів Y; N_1 та N_2 - кількість слів на вході та виході системи АРМ відповідно; N - кількість узгоджених за алгоритмом Вітербі пар вхідних та вихідних слів; H – кількість

правильно розпізнаних слів; n – обсяг словника. Обидва показники – MER та WIL – в передбачених випадках набувають значення від 0 до 1.

Недоліком показників (3.6) та (3.7) є відносна складність їх аналітичного виведення, а також складність практичної перевірки їх справедливості; також користувачу системи необхідно вміти отримати з інструментарія системи вихідні дані для проведення обчислень згідно з (3.6) [15].

3.2. Міра точності розпізнавання мови у системі автоматичного розпізнавання мови The Hidden Markov Model Toolkit

Для досягнення мети даної дисертації моделювання системи автоматичного розпізнавання мови відбувалось в програмному комплексі The Hidden Markov Model Toolkit (The HTK) – портативному програмному комплексі для побудови та використання систем, основою яких є приховані марковські моделі.

Для аналізу якості роботи системи АРМ в програмному комплексі The HTK здійснюється построкове співставлення вихідних послідовностей слів з відповідними файлами транскрипцій, засноване на алгоритмі динамічного програмування. Передбачено дві міри якості системи: *%Correct*, що враховує тільки ідентичні слова вхідних та вихідних послідовностей слів, та %Accuracy, у якому здійснюється порівняння цих же файлів на предмет наявності вставок, замін та видалень. Ці показники відповідно розраховуються за формулами (3.3) та (3.8) [8]:

$$\%Accuracy = \frac{H - I}{N} \cdot 100\% \qquad (3.8)$$

3.3. Постановка та проведення експерименту

В рамках даної дисертації було проведено чотири експерименти: навчання системи АРМ за методами Fully-Matched Training (FMT), Noise-Matched Training (NMT, інша назва Spectrum Matched Training - SMT), Signal-to-Noise Ratio Matched Training (SNRMT) та Multistyle Training (MT).

Використано 14 шумів навколишнього оточення (таблиця 3.1), що належать до різних сфер людської діяльності та найчастіше можуть бути завадами високій якості розпізнавання.

Таблиця 3.1. Шуми навколишнього оточення

Сфера застосування	Шум
Транспорт	Вулиця, вкладена бруківкою
	Площа перед вокзалом
	Тролейбусна зупинка
	Вантажівки на пр. Перемоги
Шуми приміщення: транспорт	Підземний перехід між вокзалами
	Поїзд метро під час розгону
	Фойє центрального вокзалу
	Фойє метро
	У троллейбусі
Шуми приміщення: офіс та побутові	Аудиторія
	Мікрохвильова піч
	Комп'ютер
	Кавомолка
	Пральна машина

Рівні сигнал-шум обрано з 0 до 45 дБ з кроком 5 дБ.

За методом Fully-Matched Training навчальна вибірка складалась виключно з файлів з одним спектром шуму та значенням SNR:

$$SNR_t = SNR_r$$
$$n_t(t) = n_r(t), \quad (3.9)$$

де SNR_t – відношення сигнал-завада навчальної вибірки, SNR_r – відношення сигнал-завада тестової вибірки, $n_t(t)$ – спектр шуму навчальної вибірки, $n_r(t)$ – спектр шуму тестової вибірки.

Експеримент було проведено для 140 ситуацій: по десять вибірок для кожного значення SNR для кожного з 14 шумів.

За методом Noise-Matched Training навчальна вибірка складається з сигналів, зашумленим одним певним шумом, але з усіма значеннями SNR:

$$SNR_t \neq SNR_r$$
$$n_t(t) = n_r(t) \quad (3.10)$$

В даній частині експерименту було сформовано 14 вибірок згідно з кількістю використаних шумів.

За методом Multistyle Training формується одна навчальна вибірка, яка складається з усіх можливих варіантів зашумленої мови, отже маємо різні спектри шумів та різний рівень зашумленості:

$$SNR_t \neq SNR_r$$
$$n_t(t) \neq n_r(t), \quad (3.11)$$

За методом Signal-to-Noise Ratio Matched Training навчальна вибірка складається з шумів одного і того самого рівня зашумленості, але різних спектрів:

$$SNR_t = SNR_r$$
$$n_t(t) \neq n_r(t) \quad (3.12)$$

Згідно з тим, що в рамках даної роботи відношення сигнал-завада може набувати десяти різних фіксованих значень (від 0 до 45 дБ з кроком 5 дБ), в даній частині експерименту проведено 10 дослідів.

Для наочності порівняння методів їх математичний опис також наведено в таблиці 3.2:

Таблиця 3.2. Методи навчання системи АРМ

Назва методу	*Співпадіння (Matching)*
Fully-Matched Training	$SNR_t = SNR_r; n_t(t) = n_r(t)$
Noise-Matched Training	$SNR_t \neq SNR_r;\ n_t(t) = n_r(t)$
Multistyle Training	$SNR_t \neq SNR_r;\ n_t(t) \neq n_r(t)$
Signal-to-Noise Ratio Matched Training	$SNR_t = SNR_r; n_t(t) \neq n_r(t)$

Сигнали, що зашумлюються, це попередньо записане мовлення диктора: 200 зразків 10 слів російської мови: числа від одного до десяти, кожне слово записане 20 разів з різною, наскільки це можливо, інтонацією. Фонемний словник складається з 22 фонем російської мови. SNR навчальної вибірки

становить 45 дБ. Параметри сигналів: частота дискретизації $f_d = 22050$ Гц, глибина квантування – 16 біт.

Для тестування системи АРМ була сформована адитивна суміш сигналу та шуму (рис. 3.3.1):

$$s(t) = k \cdot x(t) + n(t), \tag{3.13}$$

де $x(t)$ – мовний сигнал, $n(t)$ – шум, k – поправочний коефіцієнт, що забезпечує необхідне відношення сигнал-шум SNR_0 та розраховується за формулою:

$$k = 10^{0.05(SNR_0 - SNR)} \tag{3.14}$$

де SNR – реальне відношення сигнал-шум записаного сигналу [16].

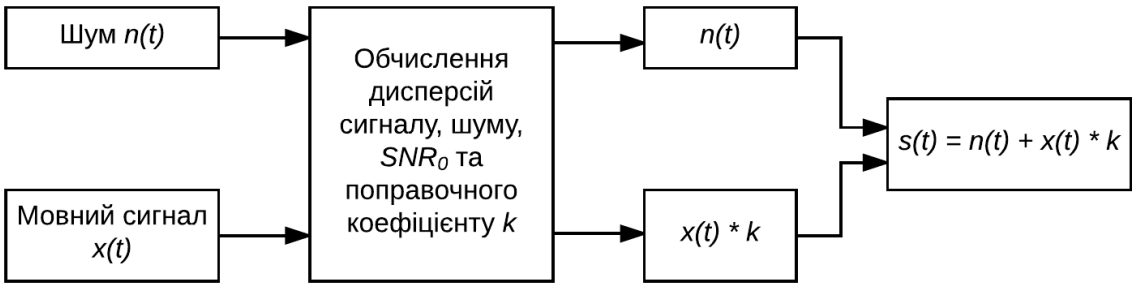

Рис. 3.3.1. Схема моделювання зашумленого сигналу

Тестові сигнали являли собою шість зашумлених звукових файлів дискретної мови із записом усіх десяти слів, що використовувались при навчанні, з паузами між словами тривалістю 0,3 – 0,5 с та різними SNR від 0 до 45 дБ.

Загальна кількість файлів, створених для вирішення задачі даної дисертації, – 28840. Для пришвидшення проведення експерименту для кожного з етапів було написано скрипти автоматизації. Лістинги програм для етапів навчання та тестування наведено в додатках А та Б відповідно.

3.4. Результати експерименту: Fully-Matched Training.

У цьому експерименті навчання системи АРМ виконувалося на зашумлених сигналах із SNR від 0 до 45 дБ з кроком 5 дб. Тестування також виконувалося на зашумлених сигналах із SNR від 0 до 45 дБ з кроком 5 дб. Метою цієї частини було встановити ефективність роботи системи, якщо навчати її на сигналах з такими самими шумовими умовами, у яких вона буде експлуатуватись. На рис. 3.4.1. відображено результати для шуму вулиці, вкладеної бруківкою; також для зручності дані представлено в таблиці 3.3.

За отриманими результатами можна зробити висновок, що найефективніше система працює у тому випадку, коли ступінь зашумленості навчальної та тестової вибірок збігається. Проте, якщо шумові умови у реальній ситуації будуть змінюватись, є ризик значної втрати якості розпізнавання [17, 18].

Таблиця 3.3. Точність розпізнавання при навчанні за методом Fully-Matched Training (шум вулиці, вкладеної бруківкою)

	Номер частини експерименту	1	2									3
	SNR тестової вибірки, дБ	SNR навчальної вибірки, дБ										
		45, «чистий» сигнал	0	5	10	15	20	25	30	35	40	45	Усі SNR
Acc, %	0	10	65	78	48	47	40	15	10	10	10	10	62
	5	12	63	90	90	85	62	50	28	15	13	12	95
	10	22	63	93	100	100	92	75	57	38	35	25	100
	15	47	47	78	95	100	100	98	90	63	65	52	100
	20	78	30	65	88	100	100	100	100	95	88	80	100
	25	88	15	58	88	97	98	100	100	100	98	97	100
	30	98	15	50	77	90	100	98	100	100	100	100	100
	35	100	13	30	72	87	95	100	100	100	100	100	100
	40	100	12	18	68	85	88	95	100	100	100	100	100
	45	100	10	12	32	50	72	85	85	92	95	97	88

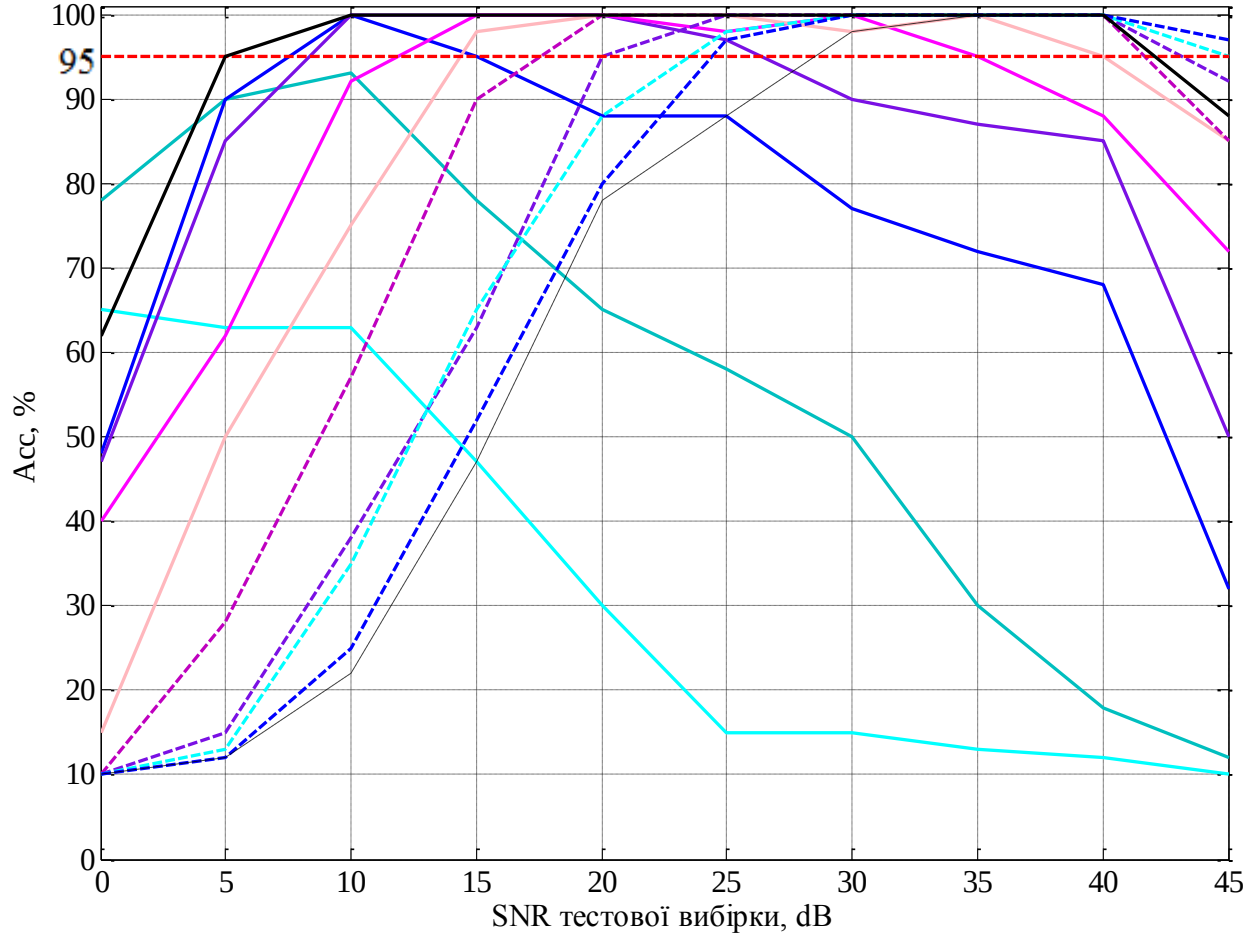

Рис. 3.4.1. Графік залежності точності розпізнавання від відношення сигнал-шум тестової вибірки для шуму вулиці, вкладеної бруківкою. SNR навчальних вибірок:

3.5. Результати експерименту: Noise Matched Training.

Метод Noise Mathed Training полягає в навчанні системи на сигналах різних способів зашумленості. Метою його було визначення того, чи стане система більш завадостійкою, якщо навчальна вибірка міститиме усі можливі варіанти шумових умов. Результати експерименту проілюстровано на рис. 3.5.1 та наведено в таблиці 3.5. [17, 18]

Таблиця 3.5. Залежність точності розпізнавання від SNR тестових вибірок, дБ

	SNR тестових вибірок, дБ	0	5	10	15	20	25	30	35	40	45
Acc, %	Вантажівки на пр. Перемоги	50	88	100	100	100	100	100	100	100	72
	Комп'ютер	50	85	97	100	100	100	100	100	100	60
	Кавомолка	58	87	100	100	98	97	95	98	98	98
	Поїзд метро під час розгону	42	45	65	80	92	97	98	100	100	85
	Мікрохвильова піч	58	87	95	100	100	100	100	100	100	80
	Привокзальна площа	72	90	93	95	95	97	100	100	100	100
	Підземний перехід між вокзалами	83	93	95	100	100	100	100	100	100	88
	Пральна машина	47	77	98	100	100	100	100	100	100	98
	Тролейбусна зупинка	60	92	97	100	100	100	100	100	100	92
	Фойє центрального вокзалу	55	93	100	100	100	100	100	100	100	97
	Вулиця, вкладена бруківкою	62	95	100	100	100	100	100	100	100	88
	Аудиторія з 13 людьми	47	67	82	87	90	95	98	98	98	80
	Фойє метро	50	88	98	100	100	100	98	98	100	85
	У троллейбусі	58	92	98	100	100	100	100	100	100	92

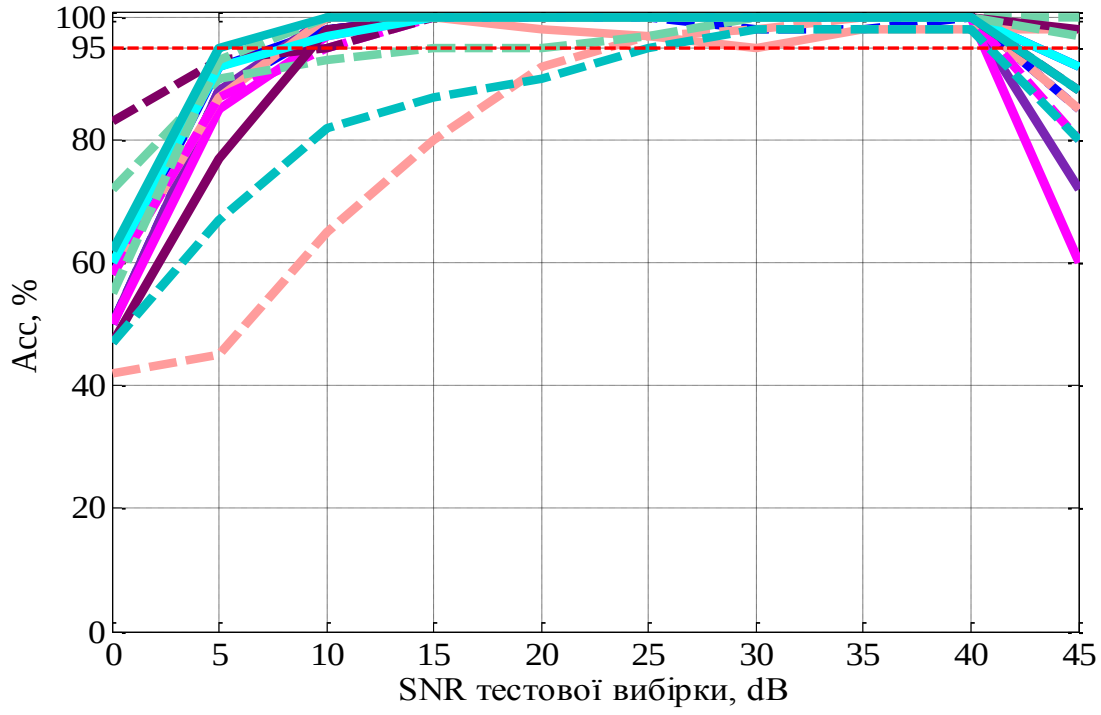

Рис. 3.5.1 Точність розпізнавання системою АРМ сигналів різної зашумленості (по горизонтальній вісі відкладені SNR тестових сигналів)

За результатами цього експерименту бачимо, що точність розпізнавання мовних сигналів суттєво збільшилась порівняно з FMT, особливо це стосується сигналів з низьким значенням SNR. У той час як при навчанні на «чистих» сигналах найкраща точність *Acc* для SNR = 3дБ дорівнювала 20% [17, 18, 19], при навчанні системи АРМ на сигналах усіх рівнів зашумленості отримуємо мінімальне значення точності *Acc* = 42%; вдалось також досягти 95% точності для сигналів з SNR = 5 дБ, а для 12 з 14 шумів точність *Acc* = 95% і вище досягається при SNR = 10 дБ, що, порівняно з навчанням на чистих, є набагато кращим результатом.

Порівнюючи дані цього етапу з даними, отриманими при навчанні за методом FMT, бачимо, що певної закономірності підвищення чи зниження якості розпізнавання немає, для різних шумів ситуація різна. Отже, при встановленні такої системи АРМ на реальні прилади чи пристрої необхідно враховувати види шумів, за яких вона буде працювати, а також можливу мінливість рівня шуму. Навчаючи систему на вибірках з різними SNR послідовно, необхідно також оснастити її програмою вимірювання відношення сигнал-шум реального сигналу для подальшого вибору необхідних еталонних зразків для розпізнавання, що займає певний обсяг пам’яті та потребує часу. У випадку навчання на вибірці, що містить усі зашумлені сигнали, цього робити не потрібно, і хоч система АРМ, навчена на вибірці такого типу, для деяких шумів показує нижчу ефективність за низьких SNR (3 дБ), вона все ж таки є більш стійкою до завад, якщо відношення сигнал-шум в процесі мовлення буде змінюватись.

Слід зазначити ще одну особливість такого навчання: під час тестування вибіркою з SNR = 45 дБ відбулась певна втрата якості розпізнавання. Пояснити це можна наступним чином: під час навчання система формує один еталонний зразок, в якому присутні спільні для усіх навчальних зразків риси. Оскільки шум в більшій чи меншій мірі вплинув на спектр більшості навчальних сигналів, еталонний зразок було створено з урахуванням цих «спотворень», і при розпізнаванні сигналів з SNR = 45 дБ система робила

помилки. Проте, у реальних умовах відношення сигнал-шум SNR ≥ 45 дБ майже не зустрічається, тож цим недоліком такого типу навчання можна знехтувати.

Порівнюючи працездатність системи АРМ в умовах різних шумів, бачимо, що найгірше розпізнавання відбувається при зашумленні шумом поїзда, що розгоняється, та шумом, записаним в аудиторії, в якій присутні 13 людей. У випадку шуму аудиторії зниження якості зумовлено відносно великою кількістю вставок (рис. 3.5.2), які виникають внаслідок того, що система помилково розпізнає шуми фонової мови як основний сигнал.

```
---------------------- Overall Results ----------------------
SENT: %Correct=0.00 [H=0, S=6, N=6]
WORD: %Corr=58.33, Acc=46.67 [H=35, D=1, S=24, I=7, N=60]
================================================================
```

Рис. 3.5.2. Скріншот результатів оцінки точності розпізнавання при тестуванні вибіркою з SNR = 0 дБ (H – кількість вірно розпізнаних слів; I – кількість вставок)

3.6. Результати експерименту: SNR-matched training.

На даному етапі було проведено 10 дослідів: кожен дослід - навчання системи на вибірці з рівнями відношення сигнал-завада SNR = 0, 5, 10… 45 дБ.

На рис. 3.4.1 наведено залежність точності розпізнавання від відношення сигнал-шум тестової вибірки для навчання на синалах з рівнем SNR = 5 дБ (синій колір). Шум тестової вибірки – фойє центрального залізничного вокзалу. Для порівняння наведено також результати навчання за методом Fully-Matched training (FMT) [17, 18]. Схожість цих двох методів полягає в тому, що рівень зашумленості співпадає, проте спектри шумів різні: виключно шум фойє центрального залізничного вокзалу для FMT та усі спектри з використаних в даній роботі для SNRMT. Як видно з графіку, метод SNRMT значно перевищує FMT по ефективності: 95% точність досягається вже при значеннях SNR>3 дБ, тоді як для методу FMT найвище значення точності сягає 70%. Слід також зазначити, що за незважаючи на високий рівень якості вхідного сигналу (SNR>30 дБ), ефективність спадає. Це пояснюється тим, що

мовні сигнали, що подаються на вхід системи при навчанні, містять додаткові спектральні компоненти, наявні у шумі. При високих значеннях SNR тестового сигналу ці компоненти відсутні, що ускладнює розпізнавання. Оскільки в реальних умовах експлуатації такі досить високі значення SNR забезпечити складно, спосіб може бути застосовним для багатьох варіантів умов, проте варто забезпечити певну стабільність рівня та спектру навколишнього шуму.

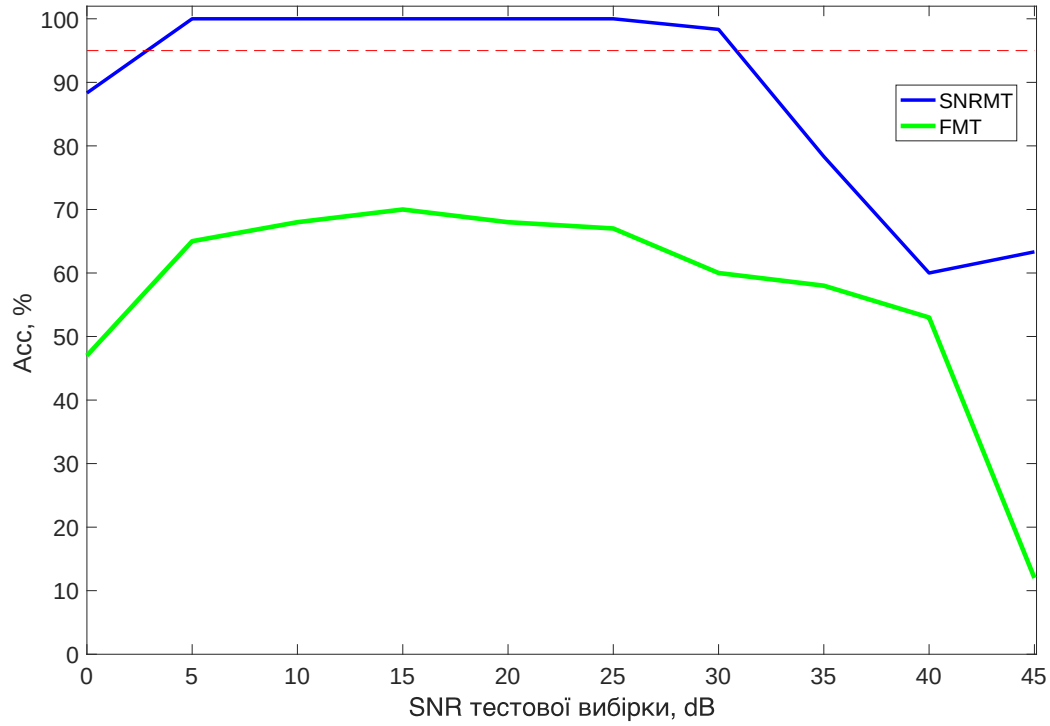

Рис. 3.4.1. Результати експерименту SNR-matched training, фойє центрального залізничного вокзалу

3.7. Результати експерименту: Multistyle Training.

В даній частині експерименту було створено одну велику навчальну вибірку з усіма можливими в рамках даної роботи варіантами зашумлення: зі значеннями відношення сигнал-шум від 0 до 45 дБ та усіма чотирнадцятьма шумами навколишнього оточення. Зашумлення відбувалось за (3.12).

Результати експерименту наведено на рис. 3.5.1 та у таблиці 3.2 [19].

Таблиця 3.2. Точність розпізнавання при навчанні методом Multistyle training

Noises	SNR_r (dB)								
	0	5	10	15	20	25	30	35	40
Вулиця, вкладена бруківкою	58	83	100	100	100	98	100	100	100
Вантажівки на пр. Перемоги	32	62	97	100	100	100	100	100	100
Тролейбусна зупинка	40	78	100	100	100	100	100	100	100
Поїзд метро під час розгону	25	50	92	97	100	100	100	100	100
Фойє метро	28	65	97	100	100	100	100	100	100
Фойє центрального вокзалу	40	78	100	100	100	100	100	100	100
Площа перед вокзалом	78	97	97	97	98	98	100	100	100
Аудиторія	28	65	97	100	100	100	100	100	100
У троллейбусі	52	87	98	100	100	98	98	100	100
Комп'ютер	18	57	93	100	100	100	100	100	100
Кавомолка	-2	15	60	88	92	93	92	92	97
Підземний перехід	83	93	95	100	100	100	100	100	100
Мікрохвильова піч	37	70	97	100	100	100	100	100	100
Пральна машина	47	72	93	100	100	100	100	100	100

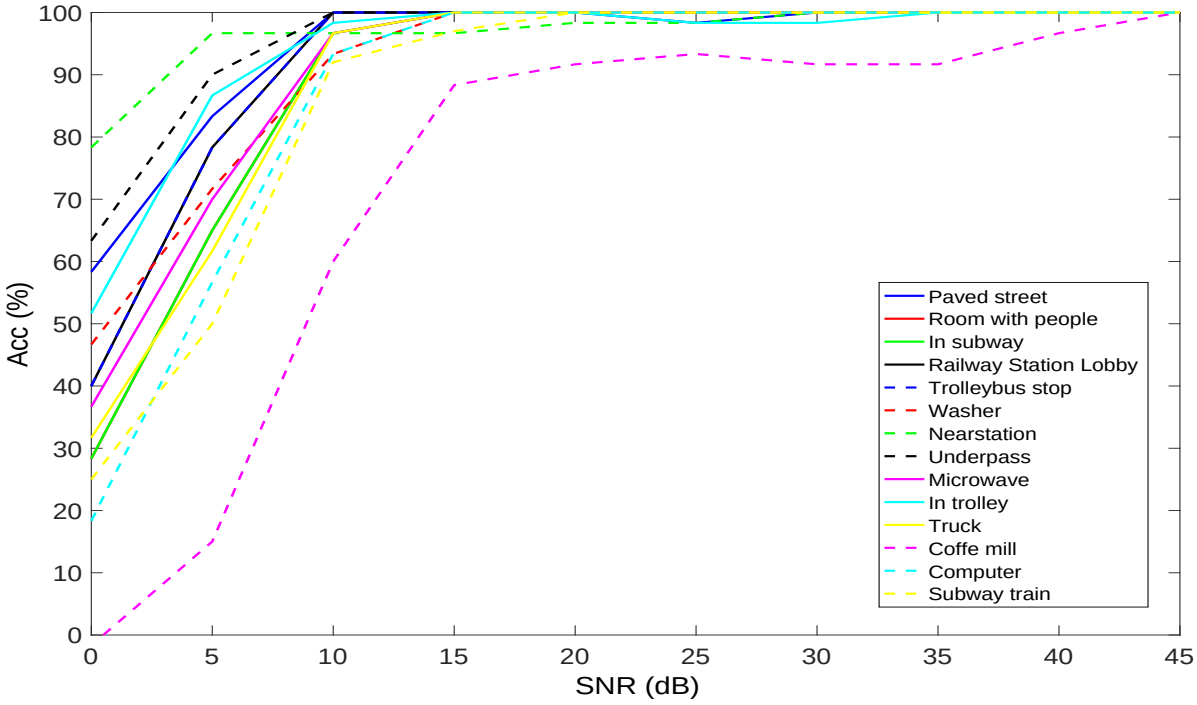

Рис. 3.5.1. Точність розпізнавання при навчанні методом Multistyle training

Бачимо, що такий варіант навчання добре підходить для умов, коли співвідношення сигнал-завада сягає або перевищує 10 дБ: тоді забезпечується дійсно висока точність розпізнавання. Перевагою такого виду навчання є рівномірність результату незалежно від спектру шуму та рівня завади: навіть якщо ми на знаємо реальних умов експлуатації, можемо підготувати систему з достатньо високим рівнем якості розпізнавання. Однак, недоліком є низькі значення точності для мовних сигналів з відношенням сигнал-завада SNR < 10 дБ. З цим можна боротися за допомогою встановлення додаткових систем шумуподавлення або ввести обмеження на умови експлуатації такої системи. [19]

3.8. Порівняння результатів та вироблення рекомендацій

Підсумовуючи усі результати, включаючи [17,18,19], можемо виробити наступні рекомендації для вибору способу навчання системи АРМ.

Навчання на суто чистих сигналах не є ефективним, воно може бути корисним виключно для отримання якісних заракетристик системи АРМ.

Навчання методом Fully Matched Training є найменш ефективним з усіх чотирьох, що порівнюються. Найвища точність забезпечується лише при співпадінні значень відношень сигнла-завада навчальної та тестової вибірок. Проте, перевагою такого способу є невелика кількість витрат ресурсів пам’яті системи: зберігається лише інформація для одного певного спектру та рівня шуму.

Навчання методом Noise Matced Training є в середньому більш ефективним, ніж методом FMT. Бачимо, що для забезпечення високої точності розпізнавання достатньо забезпечити відношення сигнал-завада вхідного сигналу не менше, ніж 10 дБ: для 12 з 14 шумів досягається точність *Acc* = 95% і вище, для одного з шумів достатньо забезпечити SNR = 5 дБ. Такий спосіб є найбільш зручним для ситуації, коли умови експлуатації не є стабільними. Певним недоліком при цьому є ресурсозатратність системи,

оскільки обсяг навчальної вибірки значно більший, а отже і інформації зберігається більше.

Навчання методом SNR-matched training доцільно порівнювати з Fully Matched Training, оскільки для них є спільним основний фактор зниження якості розпізнавання – рівень шумової завади. Вище було сказано, що SNRMT є більш точним за FMT за рахунок більшого обсягу вибірки та ширшого різноманіття спектру шумів. Висока точність розпізнавання забезпечується починаючи від значення SNR = 3 дБ, що перевищує усі наведені в даній роботі результати. Варто зазначити, що за умов нижчого рівня завади (SNR > 30 дБ) система буде втрачати у ефективності. Такі значення можуть зустрітись відносно рідко у реальних умовах експлуатації, тому навчання за методом SNRMT є одним з кращих за показниками точності, стабільності та обсягу необхідної пам’яті.

Навчання методом Multistyle training має свої переваги по відношенню до інших: починаючи з певного порогового значення SNR (від 5 до 10 дБ для різних видів шумів) система АРМ забезпечує високу точність без “провалів” при підвищенні значення SNR (як у методах FMT та SNRMT). Також очевидною перевагою є універсальність цього методу: за умов зміни спектру шуму якість розпізнавання не знизиться або знизиться незначним чином.

3.7. Висновки

При оцінюванні якості системи АРМ користуються різними мірами якості, найпоширенішими з яких є показники *WER* та *Acc*, які є простими і зручними для обчислення. Недоліком цих показників є можливість набути від’ємних значень, якщо у вихідній послідовності слів буде багато вставок. Також використовуються показники *MER* та *WIL*, які позбавлені цього недоліку, проте останній є складним для обчислення через складність вилучення з системи усієї необхідної інформації.

У системі, що використовувалась в даній роботі, – The Hidden Markov Model Toolkit – є два показники якості системи: *%Correct* та *%Acc*. У даній

роботі використовувався показник *%Acc*, оскільки *%Correct* у випадку великої кількості вставок не відображає дійсну ситуацію.

Експериментальна частина мала в собі 4 досліди згідно з кількістю розглянутих варіантів навчання системи АРМ. Розглядали способи Fully-Matched Training, Noise-Matched Training, Signal-to-Noise Ratio Matched Training та Multistyle Training. Метод Fully Matched Training виявився найменш ефективним та найменш універсальним, оскільки застосовний тільки для ситуації співпадіння значень SNR навчальної та тестово вибірок. За методом Noise Matched Training було отримано в середньому кращі результати: дослідження показало, що для забезпечення високої точності розпізнавання достатньо забезпечити значення відношення сигнал-шум не менше ніж 5 дБ. Також цей метод є більш привабливим, оскількі може бути застосовний у системах незалежно від місця іх встановлення та сфери застосування. Метод SNR-matched training дозволяє розпізнавати мову з високою точністю для сигналів, значення SNR яких перевищує 3дБ, а отже для жорстких умов є прекрасним варіантом, проте система обов'язково має працювати лище в одному типі умов експлуатації з точки зору спектру шуму. Також вартим уваги є той факт, що для тестових зразків мови зі значеннями SNR, що суттєво відрізняються від навчальних, ефективність роботи системи АРМ знижується, тобто для даного методу варто забезпечити відносну стабільність умов експлуатації. Метод Multistyle training надав кращі результати відношенню до інших: високу точність розпізнавання і для високих, і для відносно низьких значень сигнал-завада тестового сигналу. На додачу, цей метод є універсальним: за умов зміни спектру або рівня завади шуму якість розпізнавання не знизиться або знизиться незначним чином.

Також в даному розділі вироблено рекомендації по вибору методу навчання системи АРМ залежно від умов експлуатації: спектру шуму, рівня завади та мінливості цих факторів, також до уваги приймається ресурсоємність системи. Перед встановленням системи на прилад чи пристрій необхідно проаналізувати шумові умови приміщенн чи просторів, у яких

вона буде експлуатуватись, характер спектру та рівень завади на предмет мінливості рівня фонового шуму, а також необхідність залучення додаткового шумопридушуючого обладнання, оскільки від цього залежить якість роботи системи АРМ.

РОЗДІЛ 4
РОЗРОБЛЕННЯ СТАРТАП-ПРОЕКТУ

У даному розділі приведено опис розробленого стартап-проекту за темою дисертації. Наведено маркетинговий та фінансово-економічний аналіз стартап-проекту, етапи організації, заходи з комерціалізації проекту.

4.1. Мета та завдання стартап-проекту.

Впродовж останнього десятиріччя значного поширення набула форма малого венчурного підпримництва – стартап. Одними з переваг стартапів по відношенню до традиційного підприємництва є знижені бар'єри входження в ринок: завдяки доступним сучасним інструментам комунікації – Інтернет, медіапростір, соціальні мережі – значно спростився спосіб знаходження ресурсів, інвесторів, клієнтів та процес побудови та відлагодження бізнесу від ідеї до результату.

Стартап є формою венчурною, пов'язаною з великими ризиками: усього від 10% до 20% стартапів (за різними звітами) є успішними або перетворюються у стабільний бізнес. Важливими і актуальними є не тільки ідея та бачення кінцевого продукту, а також низка моментів, таких як створення бізнес-моделі, формування концепції товару, маркетинговий та фінансовий аналіз ринку, прорахунок найкращої стратегії стосовно запуску проекту, його ринкової стратегії та маркетингової програм.

На перших кроках важливо провести маркетинговий аналіз стартап проекту, тобто описати ідею проекту, визначити загальні напрями використання продукту, виокремити риси, що є переваго по відношенню до продуктів конкуруючих організацій. [20]

4.2. Опис ідеї проекту

Ідеєю даного стартап-проекту є оснащення приладів та пристроїв різного призначення системами автоматичного розпізнавання мови в якості

додаткового або основного органу керування. Оскільки результатом даної магістерської дисертації є вироблені рекомендації по вибору способу навчання систем АРМ залежно від умов експлуатації, основна проблема, що виникає під час голосового керування, є вирішеною. Пропонується наступна схема (рис.4.1):

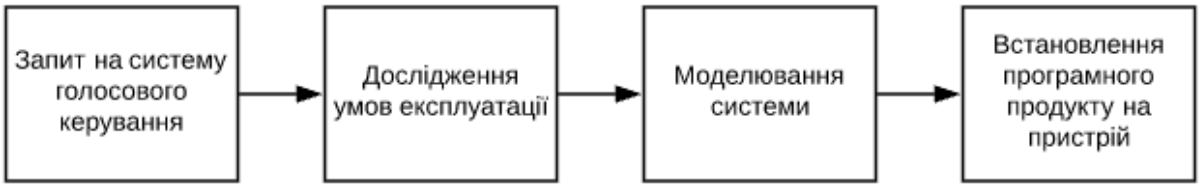

Рис. 4.1. Спрощена схема реалізації ідеї стартап-проекту

Спрощено ідею стартап-проекту можна описати наступним чином:

1. Отримання запиту на встановлення системи голосового керування на існуючий прилад або прилад, що розроблюється.
2. Дослідження умов експлуатації системи. Тут враховуються:
    a. Тип приміщення / відкрита місцевість (для врахування рівню і характеру можливих шумових завад, рівня реверберації, наявності мовленнєвих сигналів, що можуть потрапити на вхід системи).
    b. Кількість та стать дикторів (для визначення обсягу словника з точки зору тембрів голосу).
    c. Галузь використання (загальна / вузькоспеціалізована лексика, визначення обсягу словника та типу структурної одиниці).

    Також необхідно визначити ресурси процесору пристрою та спосіб імплементування системи АРМ в існуючий пристрій чи такий, що розробляється.
3. Моделювання системи відбувається з урахуванням даних, отриманих у п.2 даної схеми.
    a. Обирається спосіб навчання, за яким досягається найвища ефективність системи АРМ у заданих умовах;
    b. складається обсяг словника

с. проводиться навчання і тестування системи АРМ

4. Встановлення програмного продукту на пристрій є фінальним етапом роботи. Цей етап є одним з найважчих, оскільки кожен з пристроїв потенційних клієнтів має в основі різне апаратне та програмне забезпечення і встановлення системи АРМ та впровадження її в роботу пристрою потребує поглиблених професійних знань зі схемотехніки та програмування.

Розділ, присвячений стартап-проекту в рамках даної магістерської дисертації не потребує детального опису кожного кроку та пояснення усіх нюансів роботи, тому в рамках даної роботи технічні деталі спрощено.

За [20] опис ідеї також подається у вигляді таблиці (табл. 4.1):

Таблиця 4.1. Опис ідї стартап-проекту

Зміст ідеї	**Напрями застосування**	**Вигода для користувача**
Встановлення системи госового керування на пристрій	Наука, промисловість, криміналістика, медицина, військова справа, туризм, побут	Готове рішення поставленої задачі позбавляє замовника необхідності співпраці з декількома компаніями та розробниками; гарантія високої якості системи АРМ за рахунок теоретично та експериментально обгрунтованих рішень

У таблиці 4.2. наведено аналіз потенційних техніко-економічних переваг ідеї, визначено перелік техніко-економічних властивостей та харак- теристик ідеї, попереднє коло конкурентів (проектів-конкурентів) або товарів-замінників чи товарів-аналогів, що вже існують на ринку, та проводиться збір інформації щодо значень техніко-економічних показників для ідеї власного

проекту та проектів-конкурентів відповідно до визначеного переліку; проводено порівняльний аналіз показників. [20]

Таблиця 4.2. Аналіз потенційних техніко-економічних переваг ідеї

№ п/п	Техніко-економічні характеристики ідеї	Товари/концепції конкурентів		W (слабка сторона)	N (нейтральна сторона)	S (сильна сторона)
		Стартап-проект	CyberMova			
1.	Ціна	$150	$37,45		+	
2.	Лояльність споживачів	Висока	Висока			+
3.	Необхідність спеціальних знань для використання	Не потребується	Не потребується		+	
4.	Законодавчі обмеження	Відсутні	Відсутні			+
5.	Економія на масштабах вимірювань	Так	Так		+	
6.	Динаміка галузі	Стабільна	Стабільна		+	
7.	Інформаційне забезпечення	Добре	Посереднє			+
8.	Рівень концентрації	Низький для свого регіону	Низький для свого регіону			+
9.	Контроль якості	Проводиться	Проводиться			+
10	Кросплатформеність	Можлива	Можлива			+

4.3. Технологічний аудит

Технологічний аудит проекту полягає у визначенні технології, за якою

реалізується проект (таблиця 4.3). Визначення технологічної здійсненності ідеї проекту передбачає аналіз технології, за якою буде виготовлено товар згідно ідеї проекту, її наявність та ступінь застосовності до даного стартап-проекту та доступність. [20]

Таблиця 4.3.Технологічна здійсненність ідеї проекту

№ п/п	**Ідея проекту**	**Технології її реалізації**	**Наявність технологій**	**Доступність технологій**
1	Визначення умов експлуатації	Використання обладнання для запису	Наявні	Доступно
2	Підготовка словника	1) Запис існуючих шумів 2) Запис мови	1) Наявні 2) Наявні	1) Доступно 2) Доступно
3	Проведення процедури тестування	Програмними засобами	Наявні	Доступно
4	Імплементування програмного забезпечення або створення окрменого пристрою	Програмними та паратними засобами	Потребує доробки	Доступні

4.4. Аналіз ринкових можливостей запуску стартап-проекту

На етапі аналізу можливостей запуску стартап-проекту проводиться визначення ринкових можливостей, які можна використати під час ринкового впровадження проекту, та ринкових загроз, які можуть перешкодити реалізації проекту, дозволяє спланувати напрями розвитку проекту із урахуванням стану ринкового середовища, потреб потенційних клієнтів та пропозицій проектів-конкурентів. [20] Результати аналізу наведено в таблицях 4.4 – 4.10.

Таблиця 4.4. Характеристика ринку стартап-проекту

№ п/п	Показники стану ринку	Характеристика
1	Кількість головних гравців, од	2
2	Загальний обсяг продаж, грн/ум.од	2000
3	Динаміка ринку (якісна оцінка)	Зростає
4	Наявність обмежень для входу (вказати характер обмежень)	Відсутні
5	Специфічні вимоги до стандартизації та сертифікації	Бажано пройти державну сертифікацію у сфері стандартизації вимірювань
6	Середня норма рентабельності в галузі (або по ринку), %	65%

Таблиця 4.5. Характеристика потенційних клієнтів

№ п/п	Потреба, що формує ринок	Цільова аудиторія (цільові сегменти ринку)	Відмінності у поведінці різних потенційних цільових груп клієнтів	Вимоги споживачів до товару
	Проста та зручна система АРМ	Спеціалісти у галузях медицини, науки, криміналістики; використання у побуті та офісній роботі	Відокремленість ПО від апаратного забезпечення; необхідність спеціальної підготовки для користування системою	Швидкодія, зручність використання, зрозумілість, точність результату

Таблиця 4.6. Фактори загроз

№ п/п	Фактор	Зміст загрози	Можлива реація компанції
1	Незацікавленість аудиторії	Неготовність потенціальних клієнтів купувати продукт	Розширення можливостей комплексу, пропозиції більш дешевих варіантів
2	Поява прямого конкуренту	Втрата клієнтів, зменшення прибутків	Вдосконалення роботи у порівнянні з власним та конкурентним програмним забезпеченням
3	Постачальник бази програмного забезпечення	Заборона на реалізацію комплексу на базі цього середовища	Перехід на інше середовище

Таблиця 4.7. Ступеневий аналіз конкуренції на ринку

Особливості конкурентного середовища	В чому проявляється дана характеристика	Вплив на діяльність підприємства
Олігополія	Конкуренти зосереджені на інших регіонах світу	Підтримка результату
Світова боротьба	Відсутня локальна боротьба	Підтримка результату, захоплення підприємством стійких позицій
Внутрішньогалузева конкуренція	Спеціалісти інших галузей не можуть вплинути на ринок	Впевнена позиція в ніші
Товарно-видова конкуренція	Пропонуються замінники (наприклад, базовані на об’єктивних методах)	Створює необхідність виробництва конкурентного продукту

Таблиця 4.8. Фактори можливостей

№ п/п	Фактор	Зміст можливості	Можлива реакція компанії
1	Відсутність прямих конкурентів	Стрімкий розвиток, можливість встановлювати свою ціну	Захоплення ринку
2	Попит у інших сферах	Розширення попиту у різних сферах	Розширення ринку

Таблиця 4.9. Обгрунтування факторів конкурентоспроможності

№ п/п	Фактор конкурентоспроможності	Обгрунтування
1	Високі технічні можливості	Готове програмне та апаратне рішення
2	Конкуренти	Не існує повністю аналогічних продуктів, є постачальники окремих частин
3	Іноваційність	Даний продукт змінить ситуацію в галузі в кращу сторону за рахунок високої технологічності, новизини, та зручності

Таблиця 4.10. SWOT-аналіз проекту

Сильні сторони: Не має діючих аналогів в Україні Швидкість, зручність Готові шаблони рішень Економія ресурсів Простота використання навіть для неспеціалістів	Слабкі сторони: Необхідність адаптації ПО Створення додаткового апартаного забезпечення за необхідності
Можливості: Охоплення всіх задач акустичної експертизи Захоплення монополії на ринку	Загрози: Пропонування конкурентами кращої тех.підтримки

4.5. Розроблення ринкової стратегії проекту та маркетингової програми.

Для роботи в обраних сегментах ринку необхідно сформувати базову стратегію розвитку стартап-проекту та визначити ключові переваги потенційного продукту перед продуктами конкурентів [20]. Результати наведено в таблицях 4.11 та 4.12 відповідно.

Таблиця 4.11. Визначення базової стратегії розвитку

Обрана альтернатива розвитку проекту	**Стратегія охоплення ринку**	**Ключові конкурентоспроможні позиції відповідно до обраної альтернативи**	**Базова стратегія розвитку**
Пропонувати окремі частини комплексу	Концентрований маркетинг	Невелика кількість конкурентів, висока якість	Розширення функцій та забезпечення найвищої точності розпізнавання

Таблиця 4.12. Визначення ключових переваг концепції потенційного товару

№ п/п	**Потреба**	**Вигода, яку пропонує товар**	**Ключові переваги перед конкурентами**
1	Точність розпізнавання	Забезпечення на порядок швидшої процедури за наявні	Наявність протестованого комплексу
2	Зручність у користуванні	Пропозиція зручного інтерфейсу	Високий рівень тех.підтримки

4.6. Висновки

В даному розділі проведено аналіз запропонованого стартап-проекту: описано ідею проекту, визначено сильні, слабкі та нейтральні характеристики ідеї проекту, проведено технологічний аудит ідеї проекту (на предмет здійсненості). Проаналізовано ринкові можливості запуску стартап-проекту, охарактеризовано потенційний ринок, описано фактори загроз та

можливостей, проведено SWOT-аналіз, визначено альтернативи ринкового впровадження стартап-проекту. Розроблено ринкову стратегію проекту та маркетингову програму. За результатами аналізу можна зробити висновок, що даний стартап є інноваційним продуктом, спроможним стати сильним конкурентом існуючим фірмам та розвинутись у малий чи середній бізнес.

ВИСНОВКИ

Сучасні системи автоматичного розпізнавання мови в переважній більшості побудовано на трьох основних підходах: акусто-фонетичному, розпізнавання образів та на підході з точки зору штучного інтелекту. Також широко використовуються нейронні мережі як метод, який можна імплементувати до кожного названих підходів.

Для досягнення мети даної дисертації було використано систему, що базується на розпізнавання образів та використовує для розпізнавання приховану марковську модель, - The Hidden Markov Model Toolkit. Це система командного типу, яка добре підходить для розпізнавання дискретної мови та не потребує великих обчислювальних потужностей.

Метою експерименту було виявити вплив шумової завади на працездатність системи АРМ, а також визначити, на яких вибірках найкраще проводити навчання для досягнення найвищої якості при експлуатації у реальних шумових умовах. За результатами дослідження було зроблено наступні висновки.

Рекомендується ретельний аналіз майбутніх умов експлуатації на предмет спектру та рівню завади та імовірності її мінливості. Слід врахувати фактори, що негативно впливають на якість розпізнавання: спотворення, спричинені особливостями звукозаписувальної апаратури або спричинені реверберацією у приміщенні, та артикуляційні ефекти, та вжити заходів боротьби з ними або зниження їхнього впливу. Одним з варіантів зниження впливу несприятливих факторів є вдалий вибір способу навчання системи АРМ. Найбільш універсальними варіантами є SNR-Matched Training та Multistyle Training. SNR-Matched Training забезпечує високу точність розпізнавання для відносно низьких значень SNR вхідного сигналу, є застосовним в умовах з мінливим характером шуму, проте вразливим до зміни якості вхідного сигналу. Варіант Multistyle Training є також універсальним з точки зору застосовності при завадах різного спектру і більш стійким до зміни рівня завади. Проте, для забезпечення високої точності за низьких значень

відношення сигнал-завада, при навчанні цим способом варто оснастити систему АРМ додатковим шумопридушуючим обладнанням.

ПЕРЕЛІК ПОСИЛАНЬ

1. Дідковський В.С. Звіт за результатами НДР на тему «Дослідження шляхів підвищення ефективності слухомовної корекції людей з порушеннями слуху та глухотою / В.С. Дідковський, А.М. Продеус. – Київ, 2008.
2. Мазуренко И.Л. Компьютерные системы распознаваня речи / И.Л. Мазуренко // Интеллектуальные системы. – Москва, 1998. – № 1-2. - с.117-134.
3. Федосин С.А. Классификация систем распознаваня речи [Электронный ресурс]: / С.А. Федосин, А.Ю. Еремин. // Электронное научное издание «Электроника и информационные технологии». - 2010. – Режим доступа: http://fetmag.mrsu.ru/2010-2/pdf/SpeechRecognition.pdf
4. Rabiner, L. Fundamentals of Speech Recognition / L. Rabiner, B.H. Juang. – Prentice-Hall International, Inc. - 1993.
5. Hinton, G., Deep Neural Networks for Acoustic Modeling in Speech Recognition / G. Hinton, L. Deng, D. Yu, G.E. Dahl // IEEE Signal Processing Magazine. - November 2012. - p. 82-97.
6. Вычислительный центр им.А.А.Дороницына Российской академии наук: Модели, методы, алгоритмы и архитектуры систем распознавания речи / [отв. ред. Рязанов В.В.]. - Вычислительный центр им.А.А.Дороницына. - Москва, 2006.
7. Furtuna, T. F. Dynamic Programming Algorithms in Speech Recognition [Electronic resource] / T.F. Furtuna, // Revista Informatica Economica. - Nr. 2(46). - 2008. - p. 94-99. – Mode of access: http://revistaie.ase.ro/index.html
8. Young, S. The HTK Book (for HTK Version 3.4) / S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X.A. Liu. - Cambridge: Cambridge University Engineering Department. – 2006. – 359 p.
9. Yoshioka, T. Making Machines Understand Us in Reverberant Rooms // T. Yoshioka, A. Sehr, M. Delcroix, K. Kinoshita, R. Maas, T. Nakatani, W.

Kellerman / IEEE Signal Processing Magazine. - November 2012. - p. 114-126.

10. Lau, P. The Lombard Effect as a Communicative Phenomenon / P. Lau // UC Berkeley Phonology Lab Annual Report. – UC Berkeley. - 2008.
11. Gannot, S. Signal Enhancement Using Beamforming and Nonstationarity with Applications to Speech / S. Gannot, D. Burshtein, E. Weinstein // IEEE Transactions on Signal Processing. - August 2001. - Vol. 49 (8).
12. Juang, B.H. Speech recognition in adverse environments [Електронне джерело] / Juang B.H. // Computer Speech & Language. - July 1991. - No 5(3). - p. 275-294 .- Режим доступу: https://www.researchgate.net/publication/222192250_Speech_recognition_in_adverse_environments
13. Mansour, D. A family of distortion measures based upon projection operation for robust speech recognition / D. Mansour, B.H. Juang, // IEEE Transactions on Acoustics, Speech, and Signal Processing. - Nov, 1989. - Vol. 37, Issue 11. - p. 1659-1671.
14. Morris, A. C. From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition / V. Maier, P. Green. - Germany: Institute of Phonetics Saarland University. – 2004.
15. Kukharicheva, K. Quality Assessment of Automatic Speech recognition Systems / K. Kukharicheva // Innovations in Science and Technology: the XV All-Ukrainian R&D Students Conference Proceeding, (Kyiv, November 24, 2015) / National Technical University of Ukraine ‘Kyiv Politechnic Institute’. – Kyiv, 2015. – Part I. – 124 p., p. 109-111.
16. Дідковський В.С. Комп’ютерна обробка акустичних сигналів: Навчальний посібник / В.С. Дідковський, М.В. Дідковська, А.М. Продеус. – Київ, 2010. – 430 с.

17. Training of Automatic Speech Recognition Systems on Noised Speech / A. Prodeus, K. Kukharicheva. - 2016 IEEE 4th International Conference “Methods and Systems of Navigation and Motion Control”. – 18-20 Oct 2016 // Kyiv, Ukraine. ISBN: 978-1-5090-1052-3
18. Accuracy of Automatic Speech Recognition Systems Trained on Noised Speech / Prodeus A.N., Kukharicheva K.A. - Electronics and Control Systems №3 (49) 2016. – NAU, Kyiv, Ukraine. ISSN: 1990-5548
19. Automatic Speech Recognition Performance for Training on Noised Speech / A.Prodeus, K. Kukharicheva. - 2nd IEEE International Conference on Advanced Information and Communication Technologies Conference Proceedings. – 4-7 July, 2017// Lviv, Ukraine.
20. Розроблення стартап-проекту [Електронний ресурс]: Методичні рекомендації до виконання розділу магістерських дисертацій для студентів інженерних спеціальностей / За заг.ред. О.А.Гавриша. – Київ: НТУУ “КПІ”, 2016. – 28с.

ДОДАТОК А

Лістинг скрипту автоматизації етапу навчання системи

```
mkdir hmm1
HERest -T 3 -C config -t 250.0 150.0 1000.0 -I phones10.mlf -S train.scp -H
hmm0/hmmdefs -M hmm1 monophones
mkdir hmm2
HERest -T 3 -C config -t 250.0 150.0 1000.0 -I phones10.mlf -S train.scp -H
hmm1/hmmdefs -M hmm2 monophones
mkdir hmm3
HERest -T 3 -C config -t 250.0 150.0 1000.0 -I phones10.mlf -S train.scp -H
hmm2/hmmdefs -M hmm3 monophones
mkdir hmm4
HERest -T 3 -C config -t 250.0 150.0 1000.0 -I phones10.mlf -S train.scp -H
hmm3/hmmdefs -M hmm4 monophones
mkdir hmm5
HERest -T 3 -C config -t 250.0 150.0 1000.0 -I phones10.mlf -S train.scp -H
hmm4/hmmdefs -M hmm5 monophones
mkdir hmm6
HERest -T 3 -C config -t 250.0 150.0 1000.0 -I phones10.mlf -S train.scp -H
hmm5/hmmdefs -M hmm6 monophones
mkdir hmm7
HERest -T 3 -C config -t 250.0 150.0 1000.0 -I phones10.mlf -S train.scp -H
hmm6/hmmdefs -M hmm7 monophones
```

ДОДАТОК Б

Лістинг скрипту автоматизації етапу тестування системи

```
HParse.exe gram wdnet10
HVite -o ST -T 1 -l '*' -C config -a -H hmm1/hmmdefs -i recout1.mlf -p 0.0 -s 5.0
-S test.scp -w wdnet dict.trn monophones
HVite -o ST -T 1 -l '*' -C config -a -H hmm2/hmmdefs -i recout2.mlf -p 0.0 -s 5.0
-S test.scp -w wdnet dict.trn monophones
HVite -o ST -T 1 -l '*' -C config -a -H hmm3/hmmdefs -i recout3.mlf -p 0.0 -s 5.0
-S test.scp -w wdnet dict.trn monophones
HVite -o ST -T 1 -l '*' -C config -a -H hmm4/hmmdefs -i recout4.mlf -p 0.0 -s 5.0
-S test.scp -w wdnet dict.trn monophones
HVite -o ST -T 1 -l '*' -C config -a -H hmm5/hmmdefs -i recout5.mlf -p 0.0 -s 5.0
-S test.scp -w wdnet dict.trn monophones
HVite -o ST -T 1 -l '*' -C config -a -H hmm6/hmmdefs -i recout6.mlf -p 0.0 -s 5.0
-S test.scp -w wdnet dict.trn monophones
HVite -o ST -T 1 -l '*' -C config -a -H hmm7/hmmdefs -i recout7.mlf -p 0.0 -s 5.0
-S test.scp -w wdnet dict.trn monophones
```